nature

WHEN THE GRANTS GO AWAY

The heartache of closing a lab

PHOTOELECTRONICS
The new transistor age

HUMAN CONNECTIONS
How networks make us behave

TRADEOFFS IN DECLINE
Time for an Intervention Tax?

nature

www.nature.com/nature

Vol 457 | Issue no. 7230 | 5 February 2009

A crisis of confidence

With a surfeit of graduates for the available funds, the US scientific endeavour is increasingly losing its lustre as a career choice. The country needs to take stock and plan more carefully for the future.

Jill Rafael-Fortney and Darcy Kelley, the two scientists profiled in this issue (see page 650), are both struggling to keep their laboratories going in an extremely tough funding environment. They are also two human faces of an increasingly dire career crisis that is afflicting young researchers across the United States, as too many scientists chase an all-too-finite supply of jobs and money. The leaders of science and their allies in Congress would do well to keep such faces in mind as they map out the future of the US scientific enterprise.

The economic stimulus package currently working its way through Congress is likely to inject billions of dollars into the budgets of the scientific agencies, which will no doubt be welcome to researchers (see *Nature* **457,** 364–365; 2009). But without careful planning and sustained follow-up funding, that sudden infusion of money could end up worsening the career crisis rather then easing it. Indeed, it could readily become a replay of what happened after the budget of the National Institutes of Health (NIH) was doubled between 1998 and 2003, when the extra money drew in new scientists who then foundered when the budget virtually flat-lined in subsequent years.

The reality is that neither the United States nor any other nation knows how to calculate the number of scientists and engineers it currently needs, let alone how many it will require in the future. But at the moment, some signs suggest that the United States may have a surplus.

Supply and demand

This mismatch long predates the present financial meltdown, and it affects many areas. The career crisis is especially stark in the biomedical fields, where the number of tenure-track and tenured positions has not increased in the past two decades even as universities have nearly doubled their production of biomedical doctorates. Those who do land jobs in academic research are struggling to keep them, because competition for grant money in biomedicine has grown at a steep rate.

Some might argue that this is the way it should be. By that cool logic, the nation benefits from having a surplus of scientists who compete against each other for jobs and grants, guaranteeing that the public pays for only truly exceptional work.

There is no doubt that competition can breed excellence, and that agencies should be rigorously selective in their research portfolios. But beyond a certain point, the hyper-competition for grants, publication and tenure hurts everyone — the individuals involved, the country and science itself. The process ceases to select for only the very best young scientists, and instead starts to drive many of the smartest students out of research entirely. They realize that the risks outweigh the benefits in science and choose alternative careers. Witness the steady migration of top undergraduates to business and other professions in the past decade, and the drop in the number of doctorates in science and engineering earned by US students since it peaked in the mid-1990s. Those dedicated individuals who do stay in science find that they have less time to do the research they were trained for: a 2007 study by the Federal Demonstration Partnership in Washington DC found that investigators typically spend some 40% of their working week on grant submittals and other administrative duties.

Bursting the bubble

For researchers struggling in the current environment, we can offer no obvious solutions. But both the government and universities can do a great deal more to guard against repeating past mistakes. Most importantly, the country's leaders should heed the lessons of the NIH funding bubble and commit to long-term, predictable growth in the science budgets, not just a quick infusion of cash that will simply create another bubble (see page 649). And agencies should also refine and expand their recent efforts to direct more funding to younger investigators who bring new ideas into science.

"US leaders should commit to long-term, predictable growth in the science budgets, not just a quick infusion of cash."

At the same time, government and academic officials should look carefully at how the country trains its next crop of researchers and how many it produces — as called for by many top scientists and blue-ribbon panels. Universities, in particular, should overhaul their doctoral programmes to graduate people faster and prepare them for jobs outside the traditional academic setting.

This is made all the more urgent by the economic downturn. As in previous recessions, US research institutions are experiencing a flood of new applications for their doctoral programmes in science and engineering (see page 642). These students are entering the system at the very time when industry is cutting thousands of jobs for researchers. Hopefully those positions will reappear by the time the next wave of doctoral students graduate.

A number of resources have emerged in recent years to help young scientists start careers in and outside universities, including alternative-career clubs started by the students themselves. But these measures alone are not enough; doctoral programmes should build better career counselling and training into their curricula from the start. Reform-minded deans have long championed this change but faculty members have resisted, in part because they cling to the archaic prejudice — implicit at times — that students who leave academia are failures.

The failures, however, rest within the scientific leadership, when it focuses only on numbers and fails to see individuals who write the grant proposals, conduct the research and struggle to keep their careers afloat. ■

Against vicious activism

The US authorities need to strengthen their position on the use of animals in experiments.

Seven activists convicted of carrying out a campaign of intimidation against the animal-testing firm Huntingdon Life Sciences in Huntingdon, UK, were last month sentenced to between 4 and 11 years in prison. Hopefully, these sentences will stop future UK activists from using similar tactics, which included threats, hoax bombs, character assassination and property destruction.

Unfortunately, such tactics are increasingly being used by activists attacking scientists in California, where researchers who use animals are facing threats that include doorstep firebombs. The authorities trying to deal with this problem can find much in the UK authorities' approach to emulate.

First, activists who break the law must be vigorously pursued and prosecuted. At the same time, university leaders should set up protection plans for labs and researchers; coordinate with local and federal police before any attacks happen; and articulate a clear policy for students that legal protests are acceptable but acts of vandalism will be punished harshly.

Second, US federal, state and university authorities need to go beyond enforcement and take an unequivocal, public stand that emphasizes the importance of animal research for drug testing and basic science — as did former UK prime minister Tony Blair. It would be especially helpful if President Barack Obama were to make such a statement.

Such a level of open support might make individual researchers more apt to speak up about their own work. Britain again provides a good model in the form of Pro-Test, an activist group for those supporting animal research. Its efforts in Oxford have given a public face to supporters of animal testing.

Finally, scientists should remember that adherence to the law cuts both ways. Researchers who use animals should embrace appropriate regulations on their activities and run their labs as if members of the public could walk in at any time to take a look. If they are seen to be committed to high-quality animal care, it can only improve their credibility among the public.

Indeed, the US regulatory framework on animal research needs streamlining and strengthening. The Department of Agriculture regulates the laboratory use of cats, dogs, primates, guinea pigs and rabbits under the Animal Welfare Act, but not the ubiquitous mice and rats. It can levy fines, but tends to do so very conservatively. The Office of Laboratory Animal Welfare oversees all non-human-vertebrate research funded by the National Institutes of Health (NIH), as well as by other agencies under the purview of the NIH's parent body, the US Public Health Service. But all it can do is stop grant monies from being awarded if the institutions involved do not win its approval. Many labs also get themselves accredited by the independent Association for Assessment and Accreditation of Laboratory Care International. Its big punishment option is simply to withdraw accreditation.

The federal government should conduct a thorough review of the regulations concerning animal research to eliminate gaps, ensure compliance and strengthen penalties. Ideally, the oversight powers would be consolidated within a single organization. But, in any case, such measures might boost public confidence in animal research.

Over the long term, this multipronged approach should not only protect the safety of researchers, but should open up space for a constructive dialogue about issues in animal research — especially the pursuit of reduction, replacement and refinement of such experiments — that concern both public and researchers alike. ■

No time for rhetoric

Nicolas Sarkozy must engage with French researchers if his much-needed science reforms are to succeed.

In a speech on 22 January, as he set out his plans for a national strategy on science and innovation, French president Nicolas Sarkozy lambasted the country's university system as "infantilizing" and "paralysing for creativity and innovation". Sarkozy implied that French researchers were *fainéants* (layabouts) with cushy jobs, and no match for their supposedly more industrious British counterparts.

The speech was a typically melodramatic example of *la méthode* Sarkozy and, if it contained some home truths, it was largely a caricature. His harsh rhetoric in this case (see http://tinyurl.com/av7flg) can only reinforce the resistance he has set out to overcome. In 2000, the incumbent science minister, Claude Allègre, saw his plans for sweeping reforms dashed after scientists united against him, weary of his unnecessary provocations and sceptical of reforms imposed from on high with little consultation. Sarkozy is tempting a similar fate.

To their credit, Sarkozy and his science minister, Valérie Pécresse, have pushed through much-needed modernizations. These include putting universities on the road to independence from the centralized administration, giving them badly needed cash, and injecting a healthy dose of grants awarded on the basis of competitive proposals (see *Nature* **453,** 133; 2008).

But a massive strike across French universities that began this week (see page 640) suggests that, applied to the research community, *la méthode* Sarkozy has reached its limits. Sarkozy should heed Allègre's earlier mistakes and understand that he cannot modernize France's research system unless he has scientists on board. As things stand now, even top researchers who support the broad thrust of the reforms complain that their advice is being ignored, and that many changes seem as though they are being imposed by technocrats seeking grandiose institutional rearrangements as ends in themselves.

The substance of Sarkozy's reforms is right, but to succeed he must engage more with scientists. Many researchers experience the reforms as if they were in an aircraft flying through thick cloud, buffeted by the turbulence of almost weekly changes, with little idea of where the plane is taking them. Some fears are exaggerated, but others are legitimate. To arrive at their destination, Pécresse and Sarkozy need to consult on reforms with the navigators in the research community who know this airspace best. And Sarkozy, a speedy man, may have to accept that throttling back can sometimes avoid unwelcome accidents. ■

RESEARCH HIGHLIGHTS

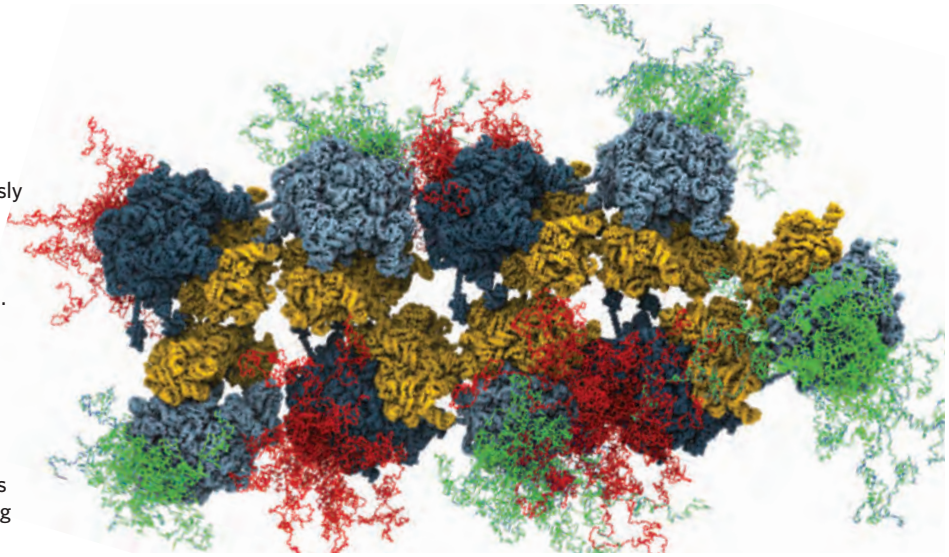
Industrial complex

Cell **136,** 261–271 (2009)

The three-dimensional structure of the polysome has been elucidated in bacteria.

Polysomes are clusters of ribosomes, the cell's protein factories. The ribosomes that make up a polysome (pictured) simultaneously read the same message, so many proteins can be made at the same time. On each ribosome the newly made protein emerges through a specialized polypeptide exit tunnel.

Ulrich Hartl and Wolfgang Baumeister at the Max Planck Institute of Biochemistry in Martinsried, Germany, and their colleagues used cryoelectron tomography to show that the ribosomes sit in either a staggered or a helical arrangement so that the exit tunnels are distant from each other. They suggest this minimizes the chance of new proteins sticking to each other.



ELSEVIER

MATERIALS SCIENCE

Graphene gets a fresh look

Phys. Rev. Lett. **102,** 026802 (2009)

The properties of graphene are tricky to understand from first principles because the material's carbon-sheet structure generates unusually strong forces between electrons. But Joaquín Drut of Ohio State University in Columbus and Timo Lähde of the University of Washington in Seattle believe they have made headway using the tools of lattice QCD, a theory from high-energy physics.

These tools allowed them to treat graphene's electrostatic interactions accurately, and, crucially, to predict that a graphene sheet should become insulating when it is not resting on another material. They hope experimentalists will soon demonstrate the insulating effect, which may have consequences for graphene-based electronics.

PHYSIOLOGY

Fake fingerprints

Science doi:10.1126/science.1166467 (2009)

Fingerprints may be important for assessing fine textures, in addition to their known role in making gripping objects easier by increasing friction.

Georges Debrégeas and his colleagues at the École Normale Supérieure in Paris made a fake fingertip and tested it with and without a print. They covered a sensor with either a smooth or ridged cap and measured pressure variations as it scanned an uneven surface. The sensor represented a nerve ending, the cap the skin.

In the experiment, ridged 'skin' amplified certain vibrations 100 times more than smooth 'skin'. The team calculated that a human fingerprint would amplify vibrations at 200–300 hertz — a range that spans those frequencies to which nerve fibres that respond to fine-texture perception are most sensitive.

SCIENCE/AAAS

MICROBIOLOGY

Community assistance

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0809533106 (2009)

A pathogen that can cause gum disease in humans uses signals from another species to bolster its defences, find Matthew Ramsey and Marvin Whiteley of the University of Texas in Austin.

They report that hydrogen peroxide secreted by *Streptococcus* bacteria stimulated *Aggregatibacter actinomycetemcomitans* to make more of a protein called ApiA, via a protein that acts as a sensor, OxyR. ApiA increases *A. actinomycetemcomitans*'s binding to a human protein called factor H, shielding this bacterial species from attack by the human immune system.

CHEMICAL BIOLOGY

Casting iron

Nature Chem. Biol. doi:10.1038/nchembio.145 (2009)

Bacteria use two biosynthetic pathways to create iron-scavenging molecules, called siderophores, that are essential to their proliferation. Most siderophore research has focused on one of these — the nonribosomal peptide synthetase (NRPS)-dependent pathway — whereas the other, the NRPS-independent siderophore (NIS) pathway, has been largely ignored.

James Naismith of the University of St Andrews in Scotland and his colleagues are the first to solve a NIS enzyme's structure: that of AcsD, which occurs in the plant pathogen *Pectobacterium chrysanthemi*.

The structure reveals that AcsD catalyses reactions between ATP, citric acid and an amino acid, L-serine, making a probable precursor to the siderophore. The authors hope their work will help in the design of inhibitors of siderophore-making enzymes from human pathogens.

ATMOSPHERIC PHYSICS

Particulate power

Geophys. Res. Lett. doi:10.1029/2008GL036350 (2009)

Aerosols emanating from vehicles and industrial plants reflect sunlight and have been linked to a cooling trend from the 1950s to the 1970s. A team led by Rolf Philipona of the Swiss federal agency MeteoSwiss has now quantified the reverse trend — warming caused by solar radiation shining through cleaner skies, which has occurred since the early 1980s.

They measured short-wave radiation,

temperature and humidity near ground level at 25 sites in Switzerland and 8 in Germany, and then derived values for long-wave radiation, including heat radiating from the ground. The data indicate that the warming attributable to incoming short-wave radiation — and thus to fewer aerosols in the atmosphere — accounts for two-thirds of the overall warming of the past two decades. This totals about 1 °C in Europe.

MOLECULAR BIOLOGY

RNA repair

Science doi:10.1126/science.1165313 (2009).
The machinery involved in the RNA interference (RNAi) pathway may protect genomes against some accidental changes in how DNA is chemically modified, geneticists have found.

Modifying DNA by adding methyl groups is a common way in which cells silence certain genes, but methylation can erode over time, making the silencing less effective.

Vincent Colot at the École Normale Supérieure in Paris and his colleagues studied *Arabidopsis thaliana* mutants that have reduced DNA methylation throughout the genome. They crossed these mutants with normal plants and found that methylation gradually returned to some genes in offspring that no longer carried the mutation. Previous research had shown that methylation doesn't return once lost. Sites that were remethylated complemented the sequence of small RNA molecules that are involved in RNAi; methylation was not restored in mutants that did not make these RNAs.

NANOTECHNOLOGY

The fine print

Nature Nanotechnol. doi:10.1038/nnano.2008.415 (2009)
The limit for information density seemed to be set: at best, a bit could be stored in the presence, or absence, of an atom or electron. But Chris Moon and his colleagues at Stanford University in California have found a way to store data in subatomic spaces using quantum holograms.

They stuck carbon monoxide molecules on a thin layer of copper using a scanning tunnelling microscope. A pond of electrons 'illuminated' the arrangement of these gas molecules, and where extra information was stored in the probabilistic shape of an electron's quantum wave, that electron formed part of a hologram. Together, the data-rich electrons formed an 'S' — for Stanford — with a linewidth as small as 0.3 nanometres.

BIOLOGY

Stench sense

J. Biol. doi:10.1186/jbiol108 (2009)
How is perception of a smell kept stable over a range of concentrations? *Drosophila melanogaster* larvae (pictured below) are attracted to the fruity odour of ethyl butyrate across a 500-fold range of concentration, according to Leslie Vosshall of the Rockefeller University in New York and her colleagues.

Mutant larvae that received input from only one of a subset of three olfactory sensory neurons were attracted to a smaller concentration range of ethyl butyrate than normal larvae, in which all three neurons were functional. Furthermore, activation of just one neuron was insufficient to trigger inhibitory neurons, which respond to high ethyl butyrate concentrations, stopping the smell from becoming overpowering.

S. GSCHMEISSNER/SPL

CHEMISTRY

Membranous mopping

Angew. Chem. Int. Edn doi:10.1002/anie.200804582 (2009)
A hollow-fibre catalytic membrane developed by researchers in China and Germany could scrub the greenhouse gas nitrous oxide from the exhausts of chemical plants.

The membrane has a type of crystal structure known as perovskite, and contains barium, cobalt, iron and zirconium. It catalyses the breakdown of nitrous oxide to free nitrogen gas and oxygen atoms, which end up bound to its surface. These atoms recombine into molecular oxygen too slowly to avoid clogging up the membrane and slowing the process. Adding methane to the system solves this problem because it mops up the oxygen as it forms. This reaction generates 'synthesis gas', a mixture commonly used in industry as a fuel or chemical feedstock.

The system's architects, Haihui Wang from South China University of Technology in Guangzhou and his colleagues, say that their membrane is the first from which oxygen can be removed quickly enough to avoid attenuating the membrane's catalytic effect.

JOURNAL CLUB

Jean Dalibard
Kastler Brossel Laboratory, CNRS, France.

A quantum-gas specialist learns about crystals from his own science.

Crystals can behave as electrical insulators or conductors. In a few crystals and under the right conditions, electrons flow perfectly. And in a subset of these superconducting crystals, the minimum temperature for perfect conduction is bizarrely warm.

On the whole, physicists have tried to explain this using models with a small number of parameters, such as the probability of an electron jumping between two sites, and the interaction energy between two neighbouring electrons. Extensive laboratory studies measuring every conceivable property of the curious crystals confirm several predictions of these models, but their general solution is still hotly debated.

Recently, a couple of research groups have been casting around for less obvious ways to understand superconducting crystals, and turned to the field that is my bread and butter: quantum gases. They have modelled electrons zooming through these crystals using gases of cold potassium atoms moving around in a space demarcated by laser beams — a kind of egg box made with light.

In December, a group led by Immanuel Bloch detected cold potassium gas switching to a state with exactly one atom per compartment of the egg box. Such an ordered state is considered a key ingredient for superconductivity. Bloch's team was not the first to see the switch, but the group's measurement of the size of the gas revealed a crucial property of this phase: its incompressibility (U. Schneider *et al. Science* **322,** 1520–1525; 2008).

This means that quantum gases are insulators as well as conductors, making the experimental analogy to superconducting crystals more complete — and making them more useful playthings for scientists studying superconducting crystals.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

NEWS

India's drug problem

Chemists show how waste-water contamination affects ecosystem.

Waste flowing out of a treatment plant near Hyderabad in India pollutes the region's waters with some of the highest levels of pharmaceuticals ever detected in the environment. In a paper being released online this week, researchers in Sweden report how this effluent has serious adverse effects on the development of tadpoles and zebrafish[1].

The findings raise concerns for the health of wildlife and ecosystems in the region, as well as underscoring little-studied potential effects on human health.

"The volume of drug production in that valley is overwhelming the system," says Stan Cox, a researcher at the Land Institute in Salina, Kansas. "Even though they have good [environmental] laws on the books, they're being swamped by the production."

"The government has not monitored for drugs being released into the environment."

For several years, the National Geophysical Research Institute in Hyderabad and the country's Central Pollution Control Board in Delhi have monitored heavy metal and other pollutants around the town of Patancheru, which is home to factories producing solvents and other chemicals. But although Patancheru is also home to numerous drug companies, the government has not monitored for drugs being released into the environment.

In 2007, however, a team led by environmental scientist Joakim Larsson of the University of Gothenburg in Sweden published results from one waste-treatment facility, Patancheru Enviro Tech Ltd (PETL)[2]. Around 90 companies in the region that manufacture active pharmaceutical ingredients, or assemble final drug products, send their waste to PETL. With permission, Larsson's team sampled the waste exiting the plant; they found drugs including the antibiotic ciprofloxacin, at concentrations of up to 31,000 micrograms per litre, and the antihistamine cetirizine, at up to 1,400 micrograms per litre. The team estimated that the amount of ciprofloxacin entering the river from the plant could amount to up to 45 kilograms a day — the equivalent of 45,000 daily doses, says Larsson.

In new work, he and co-workers exposed tadpoles and zebrafish embryos to diluted PETL effluent, equivalent to river water downstream of the plant. At the lowest concentration tested — equivalent to 1,500 cubic metres of effluent diluted in 750,000 cubic metres of river water, or a 0.2% concentration — the tadpoles experienced 40% reduced growth compared with controls. And at concentrations of 8–16%, zebrafish embryos lost colour and movement within two days of fertilization, among other developmental effects.

Larsson's team has also found drugs in

MAHESH KUMAR/AP

French scientists revolt against government reforms

University lecturers and researchers in France began a national strike on 2 February over a draft decree that would change their job descriptions and procedures for promotion.

The row has brought to a head simmering resentment among many researchers over ongoing broader reforms of research and higher education. It has been further fuelled by President Nicolas Sarkozy's criticisms of the country's researchers in a fiery speech last week.

The government's decree seems, at first glance, fairly innocuous. For the first time, evaluations of university researchers will include their contributions to teaching and university governance, and not be based solely on their research. Universities will also be given the power to change how much time staff spend on teaching and research.

So why has the decision provoked such a vocal and widespread outcry? One reason is that university researchers are used to being assessed nationally. The new policy, which is in line with the government's overall goal of giving universities greater autonomy, transfers that responsibility to the university president and board.

Scientists fear that cash-strapped universities might cut research time and force them to do more teaching, at a time when posts are being cut. In an open letter co-authored by Albert Fert, a 2007 Nobel laureate in physics from the University of Paris-Sud, top academics last week expressed worries that the changes would give university administrators too much control over scientists' work, and risk "clientship and localism".

Such concerns reflect the fact that French scientists generally trust the established

J.-P. CLATOT/AFP/GETTY IMAGES

Strikes have swept across France.

Water tested near Hyderabad contains some of the highest environmental drug levels known.

nearby lakes that do not receive effluent from the PETL plant — which suggests that drugs may also be entering the environment by means other than waste-treatment flow. Past reports, including a 2004 review commissioned by the Indian Supreme Court, noted that the PETL plant could not handle all of the waste arriving for treatment over the years. Local villagers speculated that drivers may have dumped their waste elsewhere.

The problem is gaining media attention. In January, *The Times of India* reported that the office of prime minister Manmohan Singh asked the local pollution board to start collecting data on pharmaceuticals in Patancheru's waters. And an Associated Press report last month triggered a spate of local news stories highlighting the issue.

Sri M. Narayana Reddy, president of the Hyderabad-based Bulk Drug Manufacturers Association (India), questions the validity of the research. In the past decade, Reddy says, drug manufacturers have worked to clean up their effluent, but a legacy of pollution from three decades of chemical manufacturing remains in the region's groundwater and surface water. He also notes that no manufacturer would want to lose such large quantities of a valuable drug such as ciprofloxacin. "At 20 dollars a kilogram, that's not economical," he says. "We suspect the analysis."

But within the Swedish market, Larsson's team obtained a restricted list of which companies produce or buy active pharmaceutical ingredients from India. By matching the list to records from India, they discovered that, out of 242 Swedish products studied, the active ingredient was made in India in 123 cases. Publishing online on 29 January, Larsson and Jerker Fick of Umeå University conclude that 31% of Swedish products are produced at least in part in Patancheru[3]. They propose that developed countries importing drugs should make sure that the supply chain is open, so that consumers know whether their medicines are made in an environmentally sustainable way.

The Swedish Medical Products Agency in Uppsala will lead discussions this year on how to address the country's de facto export of drug waste. The meetings will include input from the Stockholm-based Swedish Association of the Pharmaceutical Industry (LIF) as part of a special commission to review environmental impacts from manufacturing emissions nationally and internationally. "We cannot move forward on this alone," says Ethel Forsberg, director general of the Swedish Chemicals Agency, which is also party to the discussions.

She notes that polluted waters in the area are used for agriculture and also possibly for household use. The local drug manufacturers "produce medicine of very good quality," she says, "but they really cause severe damage to these people living in India around a facility like this." ■

Naomi Lubick

1. Carlsson, G., Örn, S. & Larsson, D. G. J. *Environ. Toxicol. Chem.* doi:10.1897/08-524.1 (2009).
2. Larsson, D. G. J., de Pedro, C. & Paxeus, N. *J. Hazard. Mater.* **148,** 751–755 (2007).
3. Larsson, D. G. J. & Fick, J. *Reg. Toxicol. Pharmacol.* doi:10.1016/j.yrtph.2009.01.008 (2009).

GETTY

CLEAN ENERGY CLUB
International Renewable Energy Agency founded.
www.nature.com/news

peer-review processes of the national research and higher-educational bodies, and are wary of evaluations and decisions made locally at their institutions.

In an attempt to allay these concerns, Valérie Pécresse, the minister of research and higher education, released a modified decree on 30 January that sets limits on teaching hours, and assured researchers that there would be national safeguards put in place for university promotion decisions.

Profound disarray

The spat is the first major test of the government's law on university autonomy, which was accepted with a general consensus in August 2007. Only now are the first effects of its implementation being felt. The first 20 of France's 85 universities became autonomous on 1 January 2009. They have been freed from central administrative control and are now allowed to manage their own budgets, staff and buildings, and to hire and set salaries as they see fit.

The promise of university autonomy lured Axel Kahn, a renowned geneticist at INSERM, the national biomedical research agency, to accept the presidency of the University of Paris-Descartes. Kahn, a long-standing proponent of reform, says that a major cause of researcher resentment is simply that so many reforms are being made simultaneously, prompting "profound disarray" and revolt among some scientists.

But there is also a deeply entrenched resistance among many researchers to the changing roles of key research bodies.

The large French research agencies such as the National Centre for Scientific Research (CNRS) have their own scientists and labs, and conduct most of the country's research. But Sarkozy wants to transform them into research councils, with their operational activities eventually merging with or transferring to the universities.

"I don't believe we can change any country's research system so quickly."

Many researchers fear that the government is acting too hastily, and that the university system is not ready to take on the additional research activities. "I don't believe we can change any country's research system so quickly [as the French government wants]," says one CNRS official, who requested anonymity for fear of reprisals. That's particularly true in France, he says, where most universities have been neglected for decades, and have focused on teaching large numbers of students, with most of the research being done by the agencies.

Philippe Froguel, a French scientist who heads the genomic medicine department at Imperial College London, says that he is fully in favour of plans to "responsibly transform" French universities. But, he says, apart from rare major research universities such as Kahn's, most French universities are far from ready for full autonomy. They have little experience in managing human resources and research programmes compared with the national research agencies, he says.

Kahn says that for him the right balance would be for universities to become the major operators at the local level, with research agencies maintaining their vital roles at the national and international level. "The government's vision needs to be refined a bit," he says. ■

Declan Butler

See Editorial, page 636.

The lure of the lab

Recession boosts applications to US graduate programmes.

The recession might have eroded endowments and budgets at universities around the world, but it has also brought some welcome news for higher-education institutions. Science and engineering doctoral programmes in the United States are seeing a surge in applications for the coming academic year, according to data *Nature* has obtained from several top institutions.

The boom will give universities a larger pool of potential doctoral students to choose from. And some academic leaders are encouraged by a particularly large rise in applications from domestic students, after concerns that the United States has not been producing enough home-grown scientists and engineers.

The trend fulfils expectations based on past recessions: in hard times, graduates elect to continue their education rather than take their chances on the job market. But academic leaders had wondered whether the current recession might be different if a general lack of credit prevented potential doctoral students from obtaining financial support.

The rate at which applications has risen has surprised some analysts. Debra Stewart, president of the Council of Graduate Schools in Washington DC, says that there is typically a year-long lag between the start of a recession and the application peak. She thinks that applications could continue to rise over the next year as students apply for graduate study in 2010.

The University of California, Berkeley, told *Nature* that its applications for doctoral programmes in those fields climbed by almost 7% from last year, with 11,242 people applying for programmes starting this year. Berkeley awards the greatest number of science and engineering doctorates in the United States, according to the latest comprehensive data from the National Science Foundation.

The University of Michigan in Ann Arbor, the second-largest awarder of US science and engineering doctorates, saw overall application numbers climb by 16%. Engineering applications leapt by 21%, whereas those for physics actually declined slightly. Janet Weiss, dean of Michigan's graduate school, says that applications from domestic students increased more than those from students abroad.

In recent years, some business and political leaders have expressed concern over the relative dearth of domestic students getting science and engineering doctorates. The number of such degrees granted to US citizens and permanent residents peaked in the mid-1990s, fell back again until 2002, and has climbed slowly since.

Over that same period, the number of doctorates granted to foreign students has grown

Doctoral applications are rising rapidly at the University of Michigan. T. DING/AP

Hybrid embryos fail to live up to stem-cell hopes

The creation of human–animal hybrid embryos — proposed as a way to generate embryonic stem cells without relying on scarce human eggs — has met with legislative hurdles and public outcry. But a paper published this week suggests that the approach has another, more fundamental problem: it may simply not work.

Robert Lanza of Advanced Cell Technology, a stem-cell company based in Los Angeles, California, and his colleagues show that in their labs, early-stage human–cow, human–mouse and human–rabbit hybrid embryos fail to grow beyond 16 cells (Y. Chung *et al. Cloning Stem Cells* doi:10.1089/clo.2009.0004; 2009). The hybrid embryos also failed to properly express genes thought to be critical for pluripotency — the ability to develop into a wide variety of cell types.

"It's too early to write off interspecies hybrids."

Lanza and his co-workers created their hybrid embryos using a process called somatic-cell nuclear transfer, a technique made famous when it was used to create Dolly the cloned sheep in 1996. This time, the researchers replaced the nuclei of human, cow, mouse and rabbit eggs with nuclei from human non-sex, or somatic, cells.

Human–human embryos developed normally and increased their expression of many genes, including several known to be involved in pluripotency. Hybrid embryos, however, were short-lived, and failed to express known pluripotency genes properly. Lanza says that his team has ploughed through many different protocols and "thousands" of embryos over the years, without success. "At first we thought it would just be a matter of tweaking the culture conditions," says Lanza. But "the problem was far more fundamental".

Others in the field argue that all is not lost. "Understanding what has to be done to overcome these problems would help us understand what reprogramming is all about," says reproductive biologist Justin St John of the University of Warwick in Coventry, UK, who is developing mouse–pig hybrid embryos. The paper outlines only one set of conditions used to create the embryos, St John adds, meaning that it is impossible to assess all of the options that Lanza's team tried.

Advanced Cell Technology previously cloned an endangered bull of the gaur species, using eggs from the common cow to create hybrid gaur–cow embryos. However, these two species are closely related. It may be possible to create hybrid embryos using human somatic cells and eggs from non-human primates, but primate eggs are also in short supply, says Lanza. Although Hui Zhen Sheng from the Shanghai Second Medical University in China and her colleagues have reported creating human–rabbit embryos (Y. Chen *et al. Cell Res.* 13, 251–263; 2003), several labs have been unable to replicate the findings, according to Lanza and others in the field.

There are many processes that might cause hybrid embryos to fail. Embryonic development is

by more than 70%. Although the number of foreign students dropped in the years after the 2001 terrorist attacks, it has risen in recent years.

The trend in domestic students this year was also apparent at the University of Illinois at Urbana-Champaign, another top-five school for science and engineering doctorates. Domestic applications rose by 25%, whereas foreign ones rose by 13%.

The Massachusetts Institute of Technology in Cambridge, the third-largest awarder of science and engineering doctorates, received 9,475 doctoral applications this year, a 6% increase. During the previous four years, the growth in applications had held steady at about 4% a year.

Other universities, such as the University of Toronto in Canada, the Johns Hopkins University in Baltimore, Maryland, and Duke University in Durham, North Carolina, have also reported large rises in graduate applications, says Stewart.

Although graduate deans view the numbers as a positive sign, most universities are unlikely to expand the number they enrol because they lack the resources to support more graduate students. The application surge "is happening right at the same time that most universities are undergoing the most serious financial constraints they have faced in decades and decades", says Stewart. ■

Richard Monastersky

initially guided by proteins and RNA found in the egg, with control eventually passing to DNA in the nucleus. This transfer of power occurs in humans when the embryo has reached four to eight cells; but in mice it happens at the two-cell stage, and this mismatch may disrupt development.

Furthermore, the nuclear genome may have difficulty communicating with energy-producing structures called mitochondria — which are inherited directly from the mother, through the egg — from distantly related species.

If researchers can find the reason why some hybrid embryos stop developing, they might be able to circumvent those roadblocks by altering the expression of specific genes, says cell biologist Jose Cibelli of Michigan State University in East Lansing.

Meanwhile, there may be other ways to reprogram a cell with a different species' DNA, notes embryologist Anthony Perry at the RIKEN Center for Developmental Biology in Kobe, Japan. "Is there really only one path that will give you a pup? Surely the answer is no," he says. "It's too early to write off interspecies hybrids." ■

Heidi Ledford

HAVE YOUR SAY
Comment on any of our news stories, online.
www.nature.com/news

'Experiments of concern' to be vetted online

Expert panel to offer advice on science with bioterror applications.

What do you do if you have a great idea for an experiment, but are worried that the results could enable a potential biological weapon?

Soon you will be able to ask a panel of experts for advice through a website being developed at the University of California, Berkeley.

Spearheaded by Stephen Maurer of the Goldman School of Public Policy and the Boalt Law School, the website is part of a suite of measures he has developed with scientists and public-policy experts to minimize the risks of biology research being misused.

The website, expected to begin operating by the end of March, will provide biologists with advice about 'dual-use research' or 'experiments of concern' — research with innocent goals that could inadvertently arm terrorists.

K. ANDERSON

Stephen Maurer: giving advice on tap.

Scientists will be able to enter information about proposed experiments, each of which will be reviewed by a different panel of three experts. The panels will include at least one security expert and one biologist. They will deliver a verdict on whether the work raises any security concerns, and if so, how those concerns might be addressed. The entire process should take about two weeks, says Maurer.

Maurer has lined up experiments to beta-test the site and, if those go well, the site could open for business as early as April. The portal is supported by the Carnegie Corporation of New York, a philanthropic funding body.

Experiments of concern have long troubled scientists and policy-makers, not least because most security reviews of such experiments occur at a very late stage — when the work is already finished and ready to publish.

For example, when scientists reconstructed the genome of the 1918 influenza virus and submitted their paper to *Science* in 2005, US government officials and the National Science Advisory Board for Biosecurity (NSABB) were consulted about the work. At the behest of the NSABB, *Science* ran an editorial explaining why it had published the work — even though the paper was in press by the time the advisory board made its request (see *Nature* 437, 794; 2005). The new portal is designed to provide feedback before work begins, so such problems don't arise.

"It will be a place where people can ask questions, and other people can learn from those questions, so they don't have to ask them," says Michael Imperiale, a virologist and immunologist at the University of Michigan, Ann Arbor, who is a member of the NSABB.

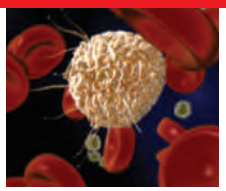
Some information about the submitted experiments will be displayed on the website, although Maurer says reasonable confidentiality measures will be undertaken, such as holding back proprietary information until the work is published.

Maurer admits that the portal's success will depend on how many scientists use it. But he is optimistic because the idea came from the community that will use it, and many scientists have already agreed to serve as expert reviewers.

"There is an instinct in the community that if you think you're talking about an experiment of concern, you should ask someone — but biosecurity people are scarce on the average campus," he says. The portal is designed to be a help, rather than a burden, in these situations. "People have enough layers of paperwork in their lives," says Maurer. "The idea is to make this as painless as possible." ■

Erika Check Hayden

Experiments of concern portal: http://tinyurl.com/bcccnu.



MEDICALRF.COM/ALAMY

STEM CELLS
Promising treatment for multiple sclerosis unveiled.
www.nature.com/news

Neanderthal genome to be unveiled

The entire genome of a 38,000-year-old Neanderthal has been sequenced by a team of scientists in Germany. The group is already extracting DNA from other ancient Neanderthal bones and hopes that the genomes will allow an unprecedented comparison between modern humans and their closest evolutionary relative.

The three-year project, which cost about €5 million (US$6.4 million), was carried out at the Max Planck Institute for Evolutionary Anthropology in Leipzig. Project leader Svante Pääbo will announce the results of the preliminary genomic analysis at the American Association for the Advancement of Science annual meeting in Chicago, Illinois, which starts on 12 February.

"We are working like crazy at the moment," says Pääbo, adding that his Max Planck colleague, computational biologist Richard Green, is coordinating the analysis of the genome's 3 billion base pairs.

Comparisons with the human genome may uncover evidence of interbreeding between Neanderthals and humans, the genomes of which overlap by more than 99%. They certainly had enough time for fraternization — *Homo sapiens* emerged as a separate species by about 400,000 years ago, and Neanderthals became extinct just 30,000 years ago. Their last common ancestor lived about 660,000 years ago, give or take 140,000 years.

The genome may also deliver more details about how these species developed their different physical traits, adapted to their environment and evolved to fight disease.

Despite previous reports that the German group's Neanderthal samples may have been contaminated with DNA from modern humans[1], an analysis of a Neanderthal mitochondrial genome[2] has allowed the researchers to largely rule out such contamination. "I have every reason to believe this is going to be authentic Neanderthal sequence," says Edward Rubin, director of the US Joint Genome Institute in Walnut Creek, California, which is also sequencing Neanderthal DNA and has collaborated with the German group in the past[3].

Almost all of the Neanderthal genome to be unveiled in Chicago comes from DNA extracted from a single bone originally discovered in a cave near Vindija in Croatia.

The age of the sample means that its DNA has degraded into fragments typically only about 50–60 base pairs long. But the German group used new sequencing technology, developed by 454 Life Sciences of Branford, Connecticut, that can analyse segments of this length.

The German team has recently extracted DNA from the bones of five other Neanderthals — and so is well on the way to creating a library of Neanderthal genomes that would allow stronger comparisons with modern humans.

Rubin's group is also sequencing Neanderthal DNA from the Croatian bone, and is trying to find other specimens to work on; as are other teams in France and Spain.

Pääbo says that his group will publish a first draft of the entire Neanderthal genome later this year, as a single read of all base pairs. However, some published human genomes had all their base pairs read eight to ten times before publication. The team says that its single-read of the Neanderthal genome is sufficient for publication because the technique used does not rely on the same DNA reassembly process used in conventional 'shotgun' sequencing. ■

Rex Dalton

1. Dalton, R. *Nature* **449,** 7 (2007).
2. Green, R. E. *et al. Cell* **134,** 416–426 (2008).
3. Noonan, J. P. *et al. Science* **314,** 1113–1118 (2006).

GRAPHIC DETAIL

Venture capital avoids bloodbath

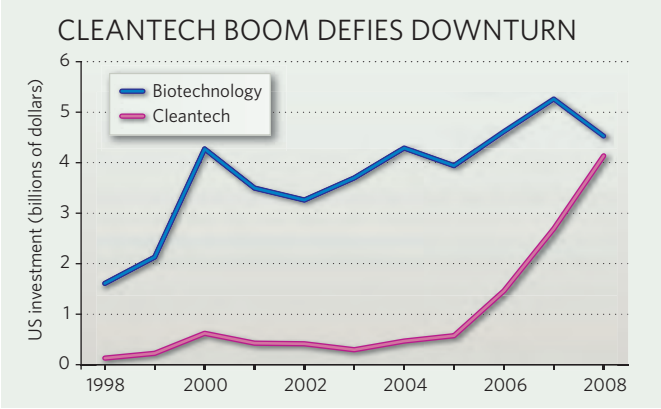
Venture capitalists raised less money and spent less money in 2008 than they did the previous year according to the MoneyTree report, a survey of venture-capital activity released on 24 January.

But not everyone emerged a loser. Investment in the fashionable clean-technology sector — encompassing everything from renewable energy to cleaner building materials — swelled by 52% last year to US$4.1 billion. "We believe that regardless of how poor the economy is, you're going to see a very strong interest in the 'cleantech' sector over the next several years," says Mark Heesen, president of the National Venture Capital Association (NVCA), based in Arlington, Virginia, which produced the report with consultants PricewaterhouseCooper.

Biotechnology did not fare as well. Investment in the industry dropped by 14% to $4.5 billion, close to the level of investment it received in 2006. In the fourth quarter — probably a better indication of how the industry will fare in 2009 — investment was 23% lower than in the fourth quarter of 2007.

Venture capitalists raised $28.0 billion for their investment funds in 2008, 21.4% less than in the previous year, and the number of mergers and acquisitions of venture-backed companies declined from 360 to 260. Only 6 venture-backed companies went public via an initial public offering, compared with 86 in 2007.

This trend may have actually helped young companies, says David Brophy of the Stephen M. Ross School of Business at the University of Michigan in Ann Arbor, because venture capitalists may turn to earlier-stage deals in the hope that the market will have recovered by the time the companies mature. Last year, the youngest of companies — 'seed' companies — received a 19% boost against the backdrop of an 8% decrease in total venture-capital investment.



Brophy says that the venture-capital community will probably experience a bigger hit in the coming year. But for some, 2008 was not as painful as expected. "Most were thinking it was going to be a bloodbath, but we didn't see that," says Heesen.

And by today's standards 'not a bloodbath' sounds refreshingly optimistic. ■

Heidi Ledford

Cash concerns for Canadian scientists

Could programme cuts prompt a brain drain?

Billions of dollars in science infrastructure investments have been overshadowed by cuts to major grant-funding programmes in Canada's federal budget.

In the Can$40-billion (US$32.3-billion) stimulus budget released on 27 January, Prime Minister Stephen Harper's government promised Can$2 billion to post-secondary institutions to repair and expand their facilities (see 'Canada's budget breakdown'). An estimated 70% of the cash will go to universities, much of it targeted towards upgrading existing labs.

University heads were delighted with the funding boost for bricks-and-mortar projects. "Infrastructure funding is something we lobbied for and is needed. It's very exciting," says Rose Goldstein, vice-president of research at the University of Calgary, Alberta.

The Canada Foundation for Innovation (CFI), which invests in research infrastructure, won an additional Can$750 million. This allows the organization to add Can$150 million to its current Can$400-million funding round, and to launch at least one additional round before the end of 2010. "Millions more dollars of really good projects are going to be funded and that's great news," says John Hepburn, vice-president of research at the University of British Columbia in Vancouver.

The budget granted Can$110 million to the Canadian Space Agency for research into space robotics over the next three years. But some of that money will be redirected from Can$9.9 million in savings the agency is supposed to generate by increasing its efficiency and effectiveness through increased collaborations with academia and industry.

Arctic research infrastructure also fared well with a Can$87-million windfall. University of Alberta ecologist David Hik says the investment is large enough to improve the ageing research stations that dot the Canadian north, but he is concerned that it is not matched by programme funding. Most Arctic research is currently supported by programmes — such as those of the International Polar Year and the Canadian Foundation for Climate and Atmospheric Sciences — that are nearing the end of their funding periods.

"The budget is good on infrastructure, but where's the money to support the graduate students, the postdocs and all the other undertakings of research that will use funding from the CFI or some other source?" says Gordon McBean, a climatologist at the University of Western Ontario in London, Ontario.

P. MCCABE/THE CANADIAN PRESS/AP PHOTO

Budget builder: Prime Minister Stephen Harper.

Although the budget does contain Can$87.5 million for graduate-student scholarships, the research community is perplexed by the government's decision to cut funding to Canada's three federal granting councils. Over three years, the budgets of the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council and the Social Sciences and Humanities Research Council will be reduced by almost Can$148 million. "It's an unfortunate consequence of getting poor advice or not listening to good advice," says Aled Edwards, a structural biologist at the University of Toronto, Ontario, and director and chief executive of the international Structural Genomics Consortium. He argues that the most efficient way to invest in research is through the funding councils, where peer review determines where the dollars are spent.

The budget made no mention of funding for Genome Canada, a not-for-profit organization that supports large-scale, multidisciplinary international science projects and regional genome centres, including the British Columbia Cancer Agency's Genome Sciences Centre in Vancouver, which completed the first publicly available draft sequence of the SARS coronavirus in 2003. The agency received five-year funding packages of Can$140 million in 2008 and Can$100 million in 2007.

"I don't think anyone understands this decision," says Hepburn. "If they forgot Can$140 million, it's really weird, and if it wasn't their intention then why isn't Martin Godbout [president and chief executive of Genome Canada] hearing any reassuring noises?" In a statement, Genome Canada noted that the lack of funding will not affect any of its current projects.

But the long-term effect of cutting funds for research may be that Canadian scientists will take their research south of the border, says Edwards. Canada's research funding pales in comparison with that in the United States, and the latest budget threatens to widen the gap between the two countries, he adds. "We're at serious risk of a brain drain." ■

Hannah Hoag

CANADA'S BUDGET BREAKDOWN

How much?	Who gets the cash?	What is the funding for?
$2 billion	Post-secondary institutions	To repair and expand facilities
$1 billion	Green Infrastructure Fund	To support projects such as sustainable energy up to 2014
$750 million	Canada Foundation for Innovation	$150 million for current funding round; $600 million to launch one or more new funding rounds by 2010
$250 million	Public Works and Government Services Canada	To modernize key federal laboratories over the next two years
$351 million	Atomic Energy of Canada	To support operations including nuclear-power research at Chalk River Laboratories, Ontario
$110 million	Canadian Space Agency	Development of prototypes for space robotics vehicles over the next three years
$87 million	Indian and Northern Affairs Canada	$2 million for High Arctic research station feasibility study; $85 million over two years to upgrade and maintain existing Arctic research facilities
$50 million	Institute for Quantum Computing (University of Waterloo, Ontario)	$25 million to complete the Quantum-Nano Centre; $25 million for operations and recruitment of researchers
$10 million	Multi-agency	To improve reporting on clean air, clean water and greenhouse-gas emissions in 2009–10

All figures in Canadian dollars.

SOURCE: BUDGET 2009: CANADA'S ECONOMIC ACTION PLAN

K. CAMPBELL/GETTY

Tighter nanotech regulations touted

The Canadian government is about to introduce the first mandatory programme in the world for reporting the safety of manufactured nanomaterials.

The scheme will require companies to provide any details that they have about the physical, chemical and toxicological properties of nanomaterials they make or import in quantities greater than one kilogram.

The government agencies Environment Canada and Health Canada will use the data to make risk assessments for the materials and to establish more specific regulations.

In 2007, the government asked the Council of Canadian Academies to assess the state of health and safety in nanotechnology. The council's panel of experts, chaired by physicist Pekka Sinervo from the University of Toronto in Ontario, reported in July 2008 that very little information existed about the risks associated with nanomaterials. "There is an urgency to come to grips with this issue," says Sinervo.

But he cautions that the success of the scheme will rely on how well the Canadian approach integrates with its main trading partners in other nations.

Voluntary data-reporting schemes have been trialled in other countries with limited success. The ongoing voluntary programme of the US Environmental Protection Agency (EPA) has so far received submissions from 29 companies on more than 120 nanoscale materials; only four companies have submitted any more than basic physical and chemical information. There has been "very little participation", says Colin Finan, from the Project on Emerging Nanotechnologies, based at the Woodrow Wilson International Center for Scholars in Washington DC.

The UK Department for Environment, Food and Rural Affairs (DEFRA) ran a two-year voluntary reporting programme from September 2006, which received 11 submissions. Across Europe, REACH (Registration, Evaluation, Authorisation and Restriction of Chemical substances) regulations are currently being reviewed to clarify how nanomaterials are dealt with.

Anne Mitchell, executive director of the Canadian Institute for Environmental Law and Policy in Toronto, expects a mandatory scheme to work more effectively than voluntary programmes, and is pleased that companies are expected to provide their data within four months of beginning to use or supply a nanomaterial.

Finan expects the United States, and perhaps other countries, to follow Canada's lead. "The EPA all but said they're going to issue regulations," he says. "Once one country starts doing something, many other countries start doing the same."

"It's going to be a useful process for the rest of the world to learn from," says Steve Morgan, nanotechnologies policy adviser at DEFRA. "We're all watching with interest." ■

Katharine Sanderson

Roche launches hostile bid for Genentech shares

Swiss pharmaceutical giant Roche made a second attempt on 30 January to capture the 44% of US biotechnology firm Genentech that it does not already own.

The board of the highly successful Genentech, based in San Francisco, California, had rejected Roche's offer of US$89 a share in summer 2008 as too low (see *Nature* 454, 381; 2008). But after the turmoil of the global financial markets saw Genentech's share value fall from a high of $99 in August last year to $84 on 29 January, Roche cut its offer to $86.50 a share and bypassed directors, going straight to shareholders.

Some analysts say that the Basel-based firm is trying to force a deal ahead of clinical-trial results expected in April that could substantially expand the use of Genentech's blockbuster anticancer drug Avastin and drive up Genentech's value. "If the trial works, Genentech is out of Roche's reach," argues Geoffrey Porges, a biotechnology analyst with Sanford Bernstein, an investment-research firm in New York.

Ebola virus hits more pig farmers in the Philippines

Four more workers at pig farms in the Philippines have contracted the Ebola-Reston subtype of Ebola virus, in addition to a case reported two weeks ago.

All five, identified by the presence of antibodies to the virus in their blood, worked with sick pigs and were probably infected more than six months ago. Ebola Reston was discovered last year in pigs on Luzon, the largest island in the Philippines (see *Nature* 457, 364–365; 2009). The virus has yet to trigger any symptoms in humans, but could mutate into more virulent forms inside pigs or other animal carriers.

R. RANOCO/REUTERS

Five people have caught the Ebola virus from pigs.

Health authorities are now testing acquaintances of the infected five to see whether human-to-human transmission might have occurred.

Refitted drilling ship sets sail

IODP-USIO

A year later than planned, the US research vessel *JOIDES Resolution* has been rebuilt. The ship departed from Singapore for Hawaii on 25 January after two years of overhaul and laboratory additions costing US$130 million (see *Nature* **453,** 7; 2008).

With more than 20 years of research-cruise experience, she is the sole US ship in the Integrated Ocean Drilling Program, which also includes the Japanese vessel *Chikyu* and European platforms. The *JOIDES Resolution* will drill cores in the bed of the equatorial Pacific to document extreme climate-change events.

Austrian scientists rattled by threat to funding

Jittery Austrian scientists have been assured that their ongoing research projects will continue to be financed, despite uncertainties over whether this year's science budget will be slashed by 40%.

The letter from Christoph Kratky, the president of the FWF, Austria's research agency, also said that no one will get a pay rise, and no new projects will be approved until the government makes its intentions clear.

Austria has been pumping up its research expenditure with special funds for several years. In 2006 it ranked fifth in research intensity among the 27 member states of the European Union. But plans to eliminate these funding pots to help fight the financial crisis were leaked from the new government when it took office in December. It is scheduled to make a final decision in April.

World's largest telescope under construction

China has begun building the biggest radio telescope in the world, the Five-hundred-meter Aperture Spherical Telescope (FAST). Sitting in a natural bowl-shaped depression in a remote region of Guizhou province, southwestern China, FAST is due to be completed in 2014.

China's National Astronomical Observatories will use the exquisite resolution of the 700-million-yuan (US$102-million) facility to identify distant pulsars and galaxies in the low-gigahertz range of the radio spectrum.

FAST will unseat the Arecibo radio telescope in Puerto Rico as the biggest single eye on the sky — although it will not be able to match the resolution of multiple-antenna telescopes such as the Very Large Array in New Mexico and the Atacama Large Millimeter/submillimeter Array (ALMA) in Chile (currently under construction).

New York tops US technology-transfer league

A survey has ranked high earners from technology-licensing income among US universities, hospitals, research institutions and investment firms during fiscal year 2007.

Of the 185 respondents that allowed their data to be published, New York University earned the most at $791 million, much of it from selling partial royalty rights to the autoimmune-disease drug Remicade. Massachusetts General Hospital in Boston was second with $346 million. Columbia University in New York, which has licensed drug therapies, medical devices and consumer electronics technologies, took in $136 million.

Research at the institutions surveyed resulted in the creation of 555 start-up companies, led by the University of California System with 38. The survey, released on 26 January, was conducted by the non-profit Association of University Technology Managers.

Correction

In the article 'China targets top talent from overseas' (*Nature* **457,** 522; 2009), we cited an incorrect salary for Shi Yigong of Tsinghua University, Beijing. *Nature* apologises for the mistake, and for any distress caused.

Beware politicians bearing gifts

The windfall for research in the proposed US stimulus package could backfire if not handled properly, warns **David Goldston**.

PARTY OF ONE

The economic stimulus package now working its way through the US Congress looks to be a bonanza for scientists. The $819-billion version written by Democrats in the House of Representatives, in concert with the administration of President Barack Obama, includes more than $13 billion in research-and-development spending. Although that figure will probably be smaller in the final bill, after negotiations with the Senate, scientists are likely to benefit immediately from the new political alignment in Washington.

The dollars slated for science are especially remarkable because they include funding for ongoing research programmes that are not usually seen as part of a stimulus effort. Indeed, science and university groups generally had lobbied only to include money to renovate laboratories because they didn't think research dollars would fit the stimulus criteria. Research grants give money to faculty members, who by definition already have a job, and to their graduate assistants, who may not even be US citizens. Research money can no doubt help the US economy now and in the longer term, but it hardly provides the same immediate boost as, say, hiring workers to build a bridge — or a laboratory, for that matter.

Not only that, the science numbers in the House bill are apparently higher even than the figures suggested privately by the new administration. This is presumably a bargaining strategy to ensure that spending does not sink too far in later compromises with the Senate.

So science advocates are right to be gleeful at this turn of events, but they should also be cautious. A stimulus bill is not the ideal vehicle for research spending, and, if scientists and their proponents aren't careful, the bill is a boon that could backfire.

First, being included in the stimulus measure could turn science spending into a political football. In general, federal support for science is something pretty much everyone in both parties agrees should be maximized, even if they haven't always followed through by providing the cash. The fight over the stimulus bill could erode that consensus, creating problems for the future. There are indications that this may already be happening. In a 24 January radio address criticizing the stimulus legislation, House Republican leader John Boehner (Ohio) complained: "There's $6 billion for colleges and universities, many of which have multibillion-dollar endowments." Interestingly, he did not describe any of this spending as 'science', perhaps fearing that might make it sound more legitimate. But then again, Obama did not mention research when describing the stimulus plan in his radio address the same day, choosing to focus on more traditional projects that would affect more Americans directly.

Second, a stimulus bill usually consists mostly of one-off projects. The idea is to inject money into the economy now that will not create long-term obligations that could swell the deficit after the economy recovers. But that's not the hope for science spending. Science proponents, both inside and outside the government, want any increases for science agencies to become part of the agencies' base spending levels, to be built on in future years. Otherwise, a brief boom could be followed by a prolonged bust, which is more or less what happened to the National Institutes of Health (NIH) earlier this decade.

But there's no guarantee that the science money will be treated differently from other stimulus spending; it's not clear whether it will end up being a down payment on future increases. The first indication of the future won't come for another month or two when Obama releases his proposed budget for fiscal 2010, which will begin on 1 October. Hedging its bets, the House did include some money for NIH grants for the next fiscal year in the stimulus legislation, but it did not do that for any other agency, and the second year of NIH funding may not make it into the final bill.

The third, and perhaps most troubling issue, is that inclusion in the stimulus bill means the science money must be awarded with unusual, perhaps even reckless, speed. With the exception of the NIH, research agencies under the House bill will have to spend the funds within 120 days. That means that the National Science Foundation (NSF), for example, would have to allocate $3 billion — a 50% increase in its budget — in four months. As of last week, the NSF was still figuring out how it could do that — whether to make more awards in whatever grant competitions it happens to have ongoing when a bill is signed, whether to revive worthy proposals from past competitions that were rejected because of lack of funds or whether to try some other strategy.

The problem is magnified for new programmes such as the Advanced Research Projects Agency-Energy, which would receive $400 million in the House bill — one-third more than Congress had previously thought the agency needed to get started. The first decisions an agency makes tend to set its course for years to come. Forcing a new agency that doesn't yet have staff to figure out how to have an impact on the nation's energy problem and award a sizable amount of money in a few months is hardly the safest way to get going (see *Nature* **447,** 130; 2007).

Moreover, spending under the stimulus bill will be under heightened scrutiny. The bill includes new watchdog and transparency provisions that will make any missteps easier to catch and more widely known. Federal science programmes are generally viewed as well managed — that's one reason they are widely supported in Washington. Any mistakes made with stimulus money may well do disproportionate damage to agencies' reputations, and mistakes are more likely with the unusual time pressure. Legislators may also press science programmes to have more immediate results because they have been designated as stimulus efforts.

None of this means that it was a mistake to include science spending in the stimulus bill. Some version of it is certain to become law, and it provides the best opportunity to reverse the stagnation in science budgets that resulted from former President George W. Bush's prolonged stalemate with Congress. But it is hardly a risk-free approach. This is a case where, even in their euphoria, scientists need to be watchful of Congress bearing gifts. ■

David Goldston is a visiting lecturer at Harvard University's Center for the Environment. Reach him at partyofonecolumn@gmail.com.

CLOSING ARGUMENTS

The battle to keep a lab funded can be long and painful. **Meredith Wadman** meets two researchers who may be close to hanging up their coats.

STONE/GETTY IMAGES

At 9:25 p.m. on Wednesday 15 October 2008, Jill Rafael-Fortney sat down at her home-office computer and wrote an e-mail to Michael Ostrowski, the chair of her department at Ohio State University in Columbus.

> Mike, I didn't get either of my grants. I just found out about the second one a few minutes ago. My career in research seems to be over. It is all I ever planned to do from the age of six, so I don't really have another well thought-out plan. Can we talk tomorrow?

Rafael-Fortney had tried, and failed, to renew the R01 grant from the National Institutes of Health (NIH) that supported her work on mouse models of muscular dystrophy. She had tried, and failed, to get a new R01 grant to study a genetic abnormality that might be widespread in human heart failure. At nearly 39, she had run out of track.

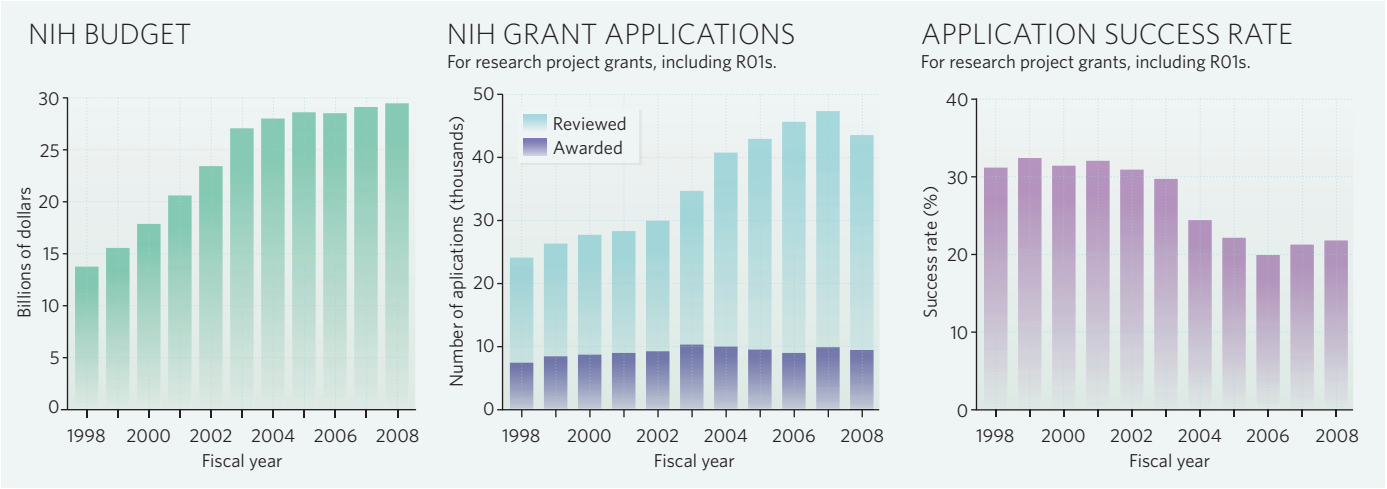
Four months earlier, at 4:00 a.m. on 10 June, Darcy Kelley had opened her laptop and logged onto the NIH's grant-review website to find out whether her own R01 application had made it through. For Kelley, a 59-year-old professor at Columbia University in New York, it was her third and final attempt to renew the major grant that supported her studies of the brain circuitry that produces and decodes sounds. She knew the results should be posted any day — and she could not sleep.

When Kelley saw that her score was 135, her heart leapt: this was outstanding on a scale in which 100 is the highest and 500 the lowest. Then as her eyes travelled further, it sank. She had heard that the National Institute of Neurological Disorders and Stroke (NINDS) was only funding proposals that scored in the 10th percentile or higher. Hers was in the 10.6th.

"You have to say that there's something wrong with the system, not the individual." — Stuart Firestein

That crushing moment of disappointment is something that countless NIH-funded scientists have shared. Between 1998 and 2003 the US Congress doubled the agency's budget to US$27.1 billion and research institutions went on a hiring boom, recruiting faculty members, postdocs and graduate students. As those scientists started and expanded labs of their own, they applied to the NIH for support. Many hoped to secure one of the three-to-five-year R01 grants that form the mainstay of biomedical research funding in the country. But just as the number of grant applications rocketed, the NIH budget flattened (see graphs, page 651). In 2000, scientists such as Rafael-Fortney and Kelley who were applying to renew a previously funded R01 had a 53% chance of success on their first submission; in 2008, according to recent NIH figures, that success rate had fallen below 24%.

There are few hard data about the types of researchers who are being squeezed out — but Rafael-Fortney and Kelley do not seem unusual. Both women run what were, even in the good times, relatively small labs. They have solid publication records, including a paper each in the *Proceedings of the National Academy of Sciences* in the past three years[1,2]. "There are literally hundreds if not thousands of people" in Rafael-Fortney's position, says Chip Souba, dean of the College of Medicine at Ohio State University. "Never in my 18–20 years of being funded by the NIH did the funding cut-off leave such a large percentage of applicants



unfunded." Stuart Firestein, one of Kelley's colleagues in Columbia's department of biological sciences, adds: "You get a score like Darcy's and you find yourself in an unfundable position, and you have to say that there's something wrong with the system, not the individual."

But others say that the system is working as it should: there has to be a line, and someone has to fall below it. "Sometimes there is a flaw in the review but usually the other proposals were just plain better," says Kathy Hudson, director of the Genetics and Public Policy Center of Johns Hopkins University in Washington DC and a former assistant director of the NIH's National Human Genome Research Institute. "We like to bemoan the limited NIH budget, and all of us who feed at the NIH trough see endless benefits to biomedical research, especially our own. But these are taxpayer dollars in hard economic times; it is not an entitlement."

Greg Simon, president of FasterCures, a group based in Washington DC that campaigns for innovation in research, says that the biomedical enterprise has outstripped the ability of the NIH and other agencies to support it, and that researchers now have to turn to private foundations and other sources of funding if they want to survive. "There has been an assumption from the way people were trained and educated that the government is in charge of full employment for research scientists," he says. "Those days are over."

Some aspects of the current NIH system have clearly placed Kelley and Rafael-Fortney at a disadvantage. Kelley would almost certainly be funded today were it not for a 2007 mandate from former NIH director Elias Zerhouni that the agency fund more investigators who had never received an R01 before. In the same funding round in which Kelley was rejected by NINDS for falling below the 10th percentile, first-time applicants needed to score only in the top 25th percentile to get through. Kelley says that the push to fund young investigators is "absolutely fair". But she also says that her situation is a "prime example of unintended consequences". She directs a highly competitive graduate neuroscience programme, and has launched dozens of successful scientists from her lab. "By supporting me, you ensure the flow of talented young investigators. If you knock me out, fewer are trained," she says.

Rafael-Fortney may have been penalized because she proposed projects that were deemed too scientifically adventurous by grant reviewers, who tend to favour studies that they know will produce the promised results. "Peer-review committees tend to be conservative and in bad times they become very conservative," Simon says. In this sense, Rafael-Fortney thinks that she is being out-competed by 'superstar' labs that boast 30 people at the bench and have the resources to collect convincing experimental data before they submit, rather than, as she did, needing the grant money to even start. Both women were also functioning on a single R01 grant in a time of diminishing NIH resources, something Kelley concedes was a mistake. "You're supposed to have two grants," she says. "It is common sense."

"Sometimes there is a flaw in the review but usually the other proposals were just plain better." — Kathy Hudson

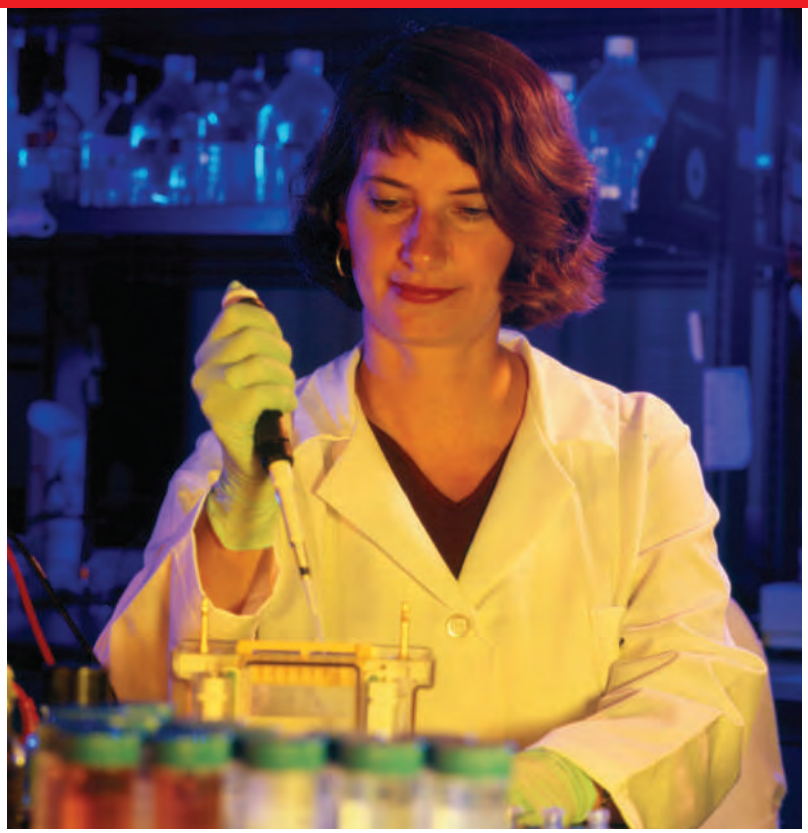
The stories of Kelley, with her long track record, and Rafael-Fortney, an investigator still trying to make her name, offer a glimpse into the personal consequences of research funding that dries up. Both now face an uncertain future — and yet both still struggle to work out exactly where they went wrong. "What I have learned from this experience is that during these times, not all good science gets funded," Rafael-Fortney says. "I don't feel like there's something that 'if only I had known' I could have done this or that." Kelley sums up her predicament more simply, saying she was "sandbagged by bad luck".

A knockout start

Late in 1996, Rafael-Fortney could be found bent over a microscope, analysing muscle tissue from a new line of transgenic mice. As a postdoctoral fellow in the lab of Kay Davies at the University of Oxford, UK, she was working on a new mouse model of muscular dystrophy, one that improved on the *mdx* mouse that was widely used at the time. The *mdx* model lacks dystrophin, the muscle protein that is missing in humans with Duchenne muscular dystrophy, but the mouse looks and acts relatively normal. A graduate student in the lab, Anne Deconinck, had knocked out a second protein called utrophin, and Rafael-Fortney found that the double knockouts had the same characteristics as patients with muscular dystrophy. "They had short stature, they were hunched up, they had difficulty breathing, they dragged their hind limbs." They also died within a few weeks.

Muscular dystrophy was what Rafael-Fortney had wanted to work on ever since, as a six-year-old, she had watched the Jerry Lewis muscular-dystrophy telethon at her home in Bayonne, New Jersey. "There were kids my age in wheelchairs," she recalls. "From that point I said: that's what I'm going do with my life. I want to find a way to help these kids and change their lives."

The mouse model, published in *Cell*[3], offered a path to such help. It became a valuable model for many muscular-dystrophy labs working to develop gene therapies against the disease. But Rafael-Fortney did not intend to be among them. "I didn't want to start a lab doing the same thing that ten other labs were doing," she says. "I really wanted to think of how else we could approach muscular dystrophy and neuromuscular diseases in general." To that end, she decided to focus on an intriguing, microscopic feature that



OHIO STATE UNIV. MEDICAL CENTER

Jill Rafael-Fortney obtained bridge funding from Ohio State University to sustain her lab.

distinguished the neurons of her double-knockout mice from those of the *mdx* mice. At the junctions where the neurons connect with muscles, one of the membranes was abnormally flat. Rafael-Fortney thought that the peculiarity might underlie a lot of their muscular-dystrophy-like symptoms, and she wanted to launch a lab to investigate.

Late in 1999, she landed a tenure-track position in the Department of Molecular and Cellular Biochemistry at Ohio State University. She got a flying start with a $400,000, three-year Burroughs Wellcome Career Award, plus $500,000 worth of awards from the American Heart Association and the Muscular Dystrophy Association. By her 31st birthday a year later she had 26 publications under her belt, ample funding and a scientific mystery to solve that completely captivated her. She knew that she needed support from the NIH as well — "it was made clear from the moment you walk in the door for a faculty position that you would not get tenure until you get an R01," she says — but this didn't seem too much of an obstacle. Congress was serving up annual increases of 15% for the biomedical agency as part of the budget doubling. Everything seemed more than on track.

The Nobel dream

At the age of 11, Kelley read a book about Nobel prizewinners and began planning her experiments. At the age of 59 she was there in Stockholm, drinking champagne and waltzing in a gold-tiled ballroom — but as a guest of her friend, department chair and Nobel prizewinner Martin Chalfie, and keenly aware that despite 108 papers, her own research career was in jeopardy.

Kelley has been at Columbia since 1982. She works mainly with frogs of the genus *Xenopus*, some 500 of which she keeps in humid, plastic tanks across the hall from her lab. *Xenopus* is valued by researchers because of its underwater vocalizations, which involve only one muscle and are not complicated by the animal's need to breathe. Kelley has published work showing that sex hormones called androgens act both in auditory brain regions and in muscle fibres in the larynx to create the differing calls of males and females[4,5]. She has also shown that an isolated larnyx[6] and a larynx connected to a brain[7] can 'sing' in a petri dish, allowing researchers to study the generation of sounds while dissecting the neural circuits involved.

In 1988 and again in 1995, Kelley won two coveted Jacob Javits Neuroscience Investigator Awards from NINDS, in which 'highly productive' scientists are selected for seven years of R01 funding. Between 1983 and 1996, Kelley also operated on a second R01 from NINDS. And she has thrown as much energy into teaching as she has into research. Fourteen years ago, she co-founded Columbia's doctoral programme in neurobiology and behaviour, which is now one of the country's most sought-after graduate programmes in neuroscience. And in 2002, she was one of 20 scientists singled out for teaching excellence by the Howard Hughes Medical Institute and awarded $1 million over four years to do innovative things with undergraduate science education. That year, Kelley had also successfully renewed her Jacob Javitz award as a standard, five-year R01. Although her second R01 had lapsed and the NIH's budget was flattening, the Howard Hughes support made her feel secure, and she used it to support undergraduate research in the lab. "I can run my lab off of one R01. And I do a bunch of other stuff. I train students. I teach. One R01 makes sense," she says.

"There are literally hundreds if not thousands of people in Rafael-Fortney's position."— Chip Souba

One R01 certainly seemed a good starting place to Rafael-Fortney in 2002 when, three years into her tenure-track job at Ohio, she got one. The National Institute of Arthritis and Musculoskeletal and Skin Diseases would pay her $910,000 over five years. The award was timely, beginning the very day after her funding from Burroughs Wellcome expired.

Because she was keen to carve out an original research path, Rafael-Fortney had proposed a relatively bold project for her first R01. The double-knockout mice had shown that an abnormal neuromuscular junction might be contributing to muscular dystrophy, and now she wanted to understand which proteins might be essential there. A large body of work pointed towards two proteins, called DLG and CASK, that have established roles in clustering channels and receptors at synapses in the central nervous system. Rafael-Fortney proposed to engineer lines of mice that would overexpress a mutated version of one of the proteins so that it no longer functioned correctly. If, as she expected, these mice had a pathology similar to muscular dystrophy, she would have started dissecting pathways with potential bearing on a broad spectrum of neuromuscular disorders.

As the results came in, though, they were disappointing. The mice didn't have muscular dystrophy. And Rafael-Fortney, who now ran a five-person lab, was starting to feel the financial pinch. By the end of 2004, her Muscular Dystrophy Association and American Heart Association grants had expired. Around two years before an R01's expiration, some investigators would begin preparing their applications for renewal, knowing that it can take many months to go through rounds of revision and resubmission. But Rafael-Fortney felt an early attempt at renewal would be "doomed" because she was still waiting for convincing experimental data. By late 2005, she was also in the late stages of her second pregnancy, and "realistically I can't function really well when I'm super pregnant", she says.

K. FRANK

After nearly 30 years of NIH support, Darcy Kelley struggled to renew her major R01 grant.

But in May 2006, on the heels of a brief maternity leave, Rafael-Fortney started working flat-out to submit her renewal. She knew that it would be tough: that year, the agency had sustained its first absolute budget cut in 36 years. "Things are looking bad," a former mentor and NIH grant reviewer warned her that summer. Rafael-Fortney was still set on bold projects though, and she was convinced that DLG and CASK were important. She proposed a project that was more technically ambitious than that in the first incarnation of the grant: knocking out DLG and CASK entirely in mouse skeletal muscle to see if their loss caused muscular dystrophy, as she still suspected it would. The thought was that removing the proteins would reveal their function more clearly than creating mutated versions. That November, one month after winning tenure, she submitted her renewal application.

Early rejection

Kelley's sole R01 came due for renewal as she was in the throes of a divorce from her husband of 33 years. It was October 2006 and her R01 was set to expire in April 2007. But after the split, she found it almost impossible to focus on the application. "My brain was blown," she says. So it was no surprise to her when, three months later, her application was 'triaged'. It had been assessed by three or more peer reviewers but then returned to her without being scored by the full 'study section' of reviewers to which it was assigned. (Between 40% and 60% of grants are routinely triaged before each study section meets.) That rejection put her at a significant disadvantage as she started re-writing the application. "When it is not scored, they don't give you the nitty gritty details of what they want you to do to improve it," says Kelley. In her application, she had proposed experiments in her frogs that would record the electrophysiological responses of single forebrain cells that both receive auditory input and direct the appropriate vocal response.

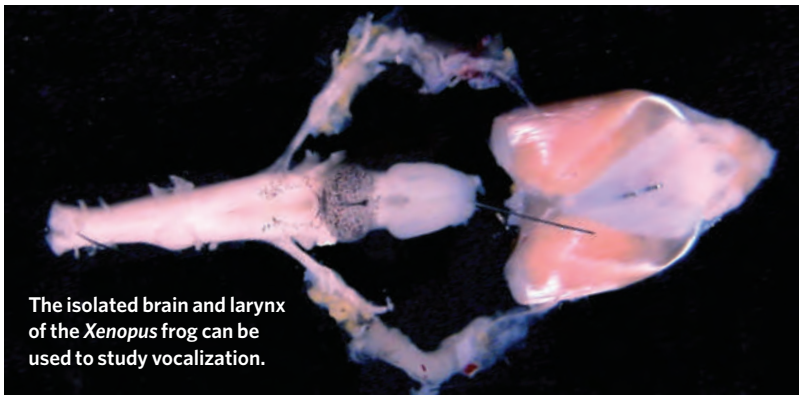
Kelley knew that factors at the NIH were working against her. Institute directors were responding to Zerhouni's Road Map for Medical Research, with its emphasis on turning basic research into clinical application. And some of those that funded animal-communication work were starting to limit applications related to model organisms to those that were directly applicable to human disease. The result was a rise in the number of animal-communication applications assigned to NINDS, even as the institute's budget was stagnating. On top of that, Zerhouni had just issued his directive that a minimum number of awards must be given to investigators who had never won an R01 before.

Grant money is more than salaries. It is also lab animals and reagents — and being able to repair the −80° freezer when it breaks down. By the time Kelley's freezer failed on New Year's eve in 2007, it was almost eight months after her R01 had expired. She had already spent months paring right back. Earlier in the year, she had told her long-standing lab technician, Candace Barnard — a single mother of two who earned less than $40,000 a year — that her position would end. And as her postdocs and graduate students finished their projects and moved on, her molecular-biology workbenches fell quiet. Their experiments cost thousands of dollars to run, whereas those recording from live frogs cost hundreds. Even here, she watched every cent. Frogs cost $25 each, so she asked her remaining students to share them when possible. Kelley also did all the surgery on the frogs, fed them, changed the water in their tanks, washed glassware and made up solutions. When her freezer — crucial for preserving years of specimens — broke, she was glad that she had hoarded the $35,000 that Columbia's administrators had given her for equipment.

Making ends meet

Rafael-Fortney was also starting to cut back. In the same month that Kelley's R01 application was triaged, Rafael-Fortney learned that hers had been, too, and it was now a certainty that she would not win renewed funding before her current R01 expired in August 2007. She applied for — and received — $120,000 in bridge funding from the university that would keep her lab afloat while she was rewriting her application. "As people left, I let them leave, to try to make remaining funds last as long as possible," she says. Rafael-Fortney's freezer — already second-hand, and stuffed with mouse tissues — gave up in October 2008. By then she could not pay for the $4,000 repair, and she was forced to store her samples in the freezers of half-a-dozen colleagues.

Throughout this time, Rafael-Fortney was writing and rewriting grant applications. The reviewers of her R01



E. ZORNIK/J. NEUROSCI.

The isolated brain and larynx of the *Xenopus* frog can be used to study vocalization.

renewal had said that the project was risky because the genetically engineered mice might not have a muscular-dystrophy phenotype. "We were asking for money to make the mouse, but they wanted it made already," she says. "It was a catch 22." But Rafael-Fortney had a second scientific string to her bow. Working with cardiac physiologist Paul Janssen from Ohio State University, she had shown that her double-knockout mice developed heart failure, which kills many patients with muscular dystrophy. One gene, called claudin 5, was expressed at strikingly low levels in the mice[8] and, she found later, in 60% of samples from human hearts that had failed for all manner of reasons[9]. William Abraham, director of the Division of Cardiovascular Medicine at Ohio, recalls being "incredibly excited" when he first learned of Rafael-Fortney's work and urging her to patent the intellectual property (which she did). Her discovery, he said, "has the potential to be a game-changer".

Rafael-Fortney and Janssen decided to apply for a 'dual PI' award, a new grant mechanism that the NIH had been promoting in which an R01 award is led by two principal investigators. They wanted to tackle two questions with the $1.25 million they hoped to share over five years: could they create heart failure in a mouse by knocking out the claudin-5 gene; and could they rescue a mouse with heart failure by inserting the gene? In February 2007 they submitted their application.

The next few months were a rollercoaster of hope and rejection. The dual PI application was returned in June, unfunded, with a score of 186 on its first assessment. That was approaching the fundable range for these applications though, and the reviewers' comments were upbeat. Hugely encouraged, Rafael-Fortney and Janssen set to work responding to the comments and shaving costs before resubmitting. Then, in February 2008, the dual PI grant was returned again, this time with a score of 213, much lower than in the previous submission. Different experts had reviewed it this time, and they had different comments, suggesting, for example, that the clinical aims of the project be removed entirely. By this time, Rafael-Fortney's R01 renewal had been returned for the second time too, unscored. Now, she had only one 'strike' — one allowable attempt — left to win funding for each grant.

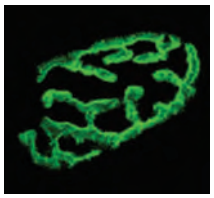
Her predicament became public in March 2008, when Rafael-Fortney was one of 12 young investigators to feature in a report entitled *A Broken Pipeline?*, compiled by leading US research institutions to plead the dire situation of young investigators. On the day of its release, Rafael-Fortney testified before the Senate Committee on Health, Education, Labor and Pensions at a hearing based on the report. "We're losing a generation of scientists," she told the senators. "They're people like me."

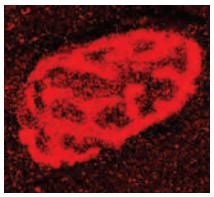
But both Rafael-Fortney and Kelley remained optimistic. Although Kelley's second renewal application had been rejected, the reviewers' comments had been positive, and she responded carefully. She submitted her third and final application on 5 March 2008. Her application was top-notch now, she felt, loaded with preliminary data and ripe for funding.

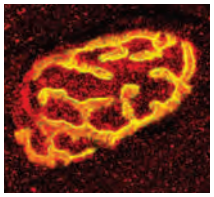
Rafael-Fortney, too, thought that her third application was convincing. With the help of another lab, she was able to generate a mouse line with CASK gene expression selectively knocked out in its skeletal muscle. Just days before the last application was due, she analysed the first of these mice. Their pathology showed that they did indeed have muscular dystrophy. She gave this top billing in her final application — she had completed a $30,000 piece of the proposed project, and had shown that the resulting mouse was a prime target for studying the disease. In effect, the application said "Look, we made the mice. We are really on to something," she says.

In the summer of 2008, Rafael-Fortney deferred a family vacation because there was no longer grant money to pay her summer salary. She submitted her final applications for both grants. Then she waited.

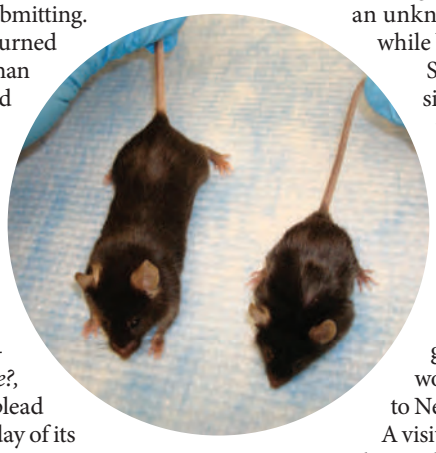
J. L. SANFORD







J. L. SANFORD



Mouse models of muscular dystrophy (bottom) and studies of the neuromuscular junction (top) could open up new angles for research on the disease.

Desperate measures

On 10 June, the morning that Kelley learned that her third and final R01 application had failed, she started making phone calls. Merrill Mitler, the programme officer who was her main staff contact at NINDS replied by e-mail: the funding cut-off was "holding firm at the 10th percentile", he wrote, and her options there were limited. Kelley started chasing down Lana Shekim, the director of voice and speech programmes at the NIH's National Institute on Deafness and Other Communication Disorders. When she had submitted her renewal, Kelley had requested that the institute, which was funding down to the 21st percentile, consider funding her grant should it fail to make the cut at the hard-pressed neurology institute.

Shekim and Kelley finally connected by phone on 27 June. Kelley was in her car, returning to New York after giving a lecture at the Marine Biological Laboratory in Woods Hole, Massachusetts. She pulled over in a gravel parking lot, thinking it would be unwise to try to persuade an unknown bureaucrat to sustain her scientific future while behind the wheel.

Shekim explained that Kelley's work was too far outside the institute's normal portfolio, which limited the animal work it funded to models of specific human disorders. "The science in this proposal does not directly address voice and speech disorders," Shekim later told *Nature*. Kelley disagreed — arguing that *Xenopus* can reveal conserved mechanisms by which nerve circuits generate vocal patterns. "I yelled at Lana. I said 'Look, I'm fighting for my scientific life here. I don't think you guys want to knock me out of science.'" Finally, a worn-out Kelley hung up and finished the long drive to New York.

A visitor to Kelley's lab last November would have found almost all her workbenches empty. One of her two remaining graduate students — whose $28,000 annual stipend the university had agreed to start paying — was preparing a list of needed supplies. A colleague of Kelley's was quietly paying for the supplies: the lab was effectively broke. Kelley knows now that she should have kept a second supporting grant in place. "The NIH funding programme takes people like me who run small labs, and have only a single grant, and penalizes them basically," she says. "Because they are not playing the game. They are not playing this multi-grant game." But others say that researchers such as Kelley need to anticipate how the funding environment will change and adjust their strategy accordingly. "If the NIH is turning towards a more

K. FRANK

clinically focused and translational-medicine emphasis, then it is researchers' responsibility to figure out how their research can fit in," says Hudson.

There is still hope for Kelley. NINDS' advisory council will meet this month, and it can make an executive decision to reach below its 10th-percentile cut-off and fund her grant. Kelley will be anxiously awaiting a post-meeting call from Mitler, who will be in the room. And even if the answer is no, Kelley says she is far from giving up. "I don't seem to be crushed. Nothing is going to keep me from doing science," she says. Kelley says that if she could go back and talk to herself as an 11 year old, reading about Nobel prize-winners, she would give nothing but encouragement. "Even if you have to go through this horrible struggle," she says, "it's so much fun that there's just no imagining anything else you could do."

Cutting costs

On 17 October 2008, Rafael-Fortney was in the midst of a grim task: killing two lines of her knockout mice. Two days earlier, she had heard that both her last-chance R01 applications had been returned unscored. Now in her 14th month of bridge funding, the $1,000 in monthly maintenance costs for the mice was no longer tenable. Some 100 of them had to be dispatched.

K. FRANK

Quiet times: Darcy Kelley, and her one senior research scientist, look after the frogs.

The comments on both her R01 rejections had been hard to swallow. Despite having added preliminary results for her CASK knockout mouse, the reviewers still complained that she didn't have the other mouse — the DLG knockout — in hand. And the reviewers on the dual PI grant considered the bid to make a claudin-5 knockout a risk too, even though, in earlier rounds, they had expressed great confidence in her ability to do so. Part of the problem, Rafael-Fortney thinks, is that she occupies a difficult middle ground in her scientific career: not senior enough to have years of resources and networks to fall back on in hard times, nor 'young' enough — ten years or closer to her PhD — to profit from NIH programmes targeted at the newest scientists. "Mostly the superstars are getting funded," she says. One NIH administrator had told her: "Unfortunately you're not a giant in the field and you're not at Harvard."

After receiving Rafael-Fortney's desperate e-mail and talking to her at length, her department chair Ostrowski helped her regroup. She has submitted new grants to the American Heart Association and the Muscular Dystrophy Association. At the same time, she has rewritten the failed R01 grant application, stripping it of half of its experiments and adding new ones to qualify it as an entirely fresh application, this time focused on the role of CASK in developing adult skeletal muscle. Events on the national stage could now work in her favour. President Barack Obama's budget request for 2010, due for release in March, and a 2009 budget now being finalized by Congress, could provide a new influx of funds for the NIH. And the economic stimulus bill being discussed by US lawmakers could provide more. But this time around, Rafael-Fortney will only have two chances to apply: in January 2009, the NIH implemented rules that R01 applications can only be submitted twice.

But Rafael-Fortney says that she can't give up. "It would mean giving up what I was passionate about, what I have been passionate about my whole life," she says. And she is also unwilling to compromise her scientific aims by proposing less ambitious projects. Recently, she finished reviewing a stack of fellowship applications for the NIH and she says she was tempted to favour the 'safer' proposal, the one with all the preliminary data in hand. "It's really hard to not go for the one experiment that's almost done," she says. Her fear is that if she starts rejecting risky projects, as hers were rejected, then more researchers in her position will be shut out. "It's going to just wipe out the whole middle," she says. "Everyone between the ages of 35 and 50 are just going to be gone in science." ■

"You're supposed to have two grants. It's common sense." — Darcy Kelley

Meredith Wadman is a reporter for *Nature* based in Washington DC.

1. Hanft, L. M., Rybakova, I. N., Patel, J. R., Rafael-Fortney, J. A. & Ervasti, J. M. *Proc. Natl Acad. Sci. USA* **103,** 5385–5390 (2006).
2. Yang, E.-J., Nasipak, B. T. & Kelley, D. B. *Proc. Natl Acad. Sci. USA* **104,** 2477–2482 (2007).
3. Deconinck, A. E. *et al. Cell* **90,** 717–727 (1997).
4. Kelley, D. B. *Science* **207,** 553–555 (1980).
5. Sassoon, D. A, Gray, G. E. & Kelley, D. B. *J. Neurosci.* **7,** 3198–3206 (1987).
6. Tobias, M. L. & Kelley, D. B. *J. Neurosci.* **7,** 3191–3197 (1987).
7. Zornik, E. & Kelley, D. B. *J. Neurosci.* **28,** 612–621 (2008).
8. Sanford, J. L. *et al. J. Mol. Cell. Cardiol.* **38,** 323–332 (2005).
9. Mays, T. A. *et al. J. Mol. Cell. Cardiol.* **45,** 81–87(2008).

See Editorial, page 635.

CORRESPONDENCE

Arizona's big city lights are damaging astronomy

SIR — As Malcolm Smith points out in his Commentary 'Time to turn off the lights' (*Nature* **457,** 27; 2009), significant economic benefits are to be had in the form of reduced electricity bills once efficient, astronomy-friendly outdoor lighting is adopted.

In locations such as Arizona, Hawaii and Chile, which are particularly suited to ground-based observational astronomy, the failure of governments to enact and enforce appropriate light-pollution controls could eventually lead to great economic losses. A study conducted by the University of Arizona's Eller College of Management, and sponsored by the Arizona Arts, Sciences and Technology Academy, showed that in financial year 2005–06 the total economic impact in Arizona of astronomical and space-science research was over US$250 million. More than 3,300 jobs were supported directly or indirectly by the flow of these dollars into the state (www.aasta.net). Although I am not aware of comparable studies of the economic impact of astronomy and space science in Hawaii and Chile, it is evidently significant.

Cities in Arizona, such as Flagstaff and Tucson, and the surrounding Coconino and Pima counties, long ago enacted lighting ordinances designed to protect the nearby observatories, and these are enforced. However, our dark skies continue to be degraded by light pollution originating from more distant, larger and rapidly growing regions, such as the greater Phoenix metropolitan area and Pinal County between Phoenix and Tucson, where lighting ordinances are typically less stringent. Unless better lighting controls are enacted in these areas, the competitiveness of Arizona's observatories will be harmed and the state, instead of benefiting from further growth of this clean and green enterprise, could experience a serious decline in astronomy's contribution to the state's economy.

I do not know whether observatories in Hawaii and Chile are vulnerable to a threat of similar magnitude, but I very much doubt they are immune.

Robert L. Millis Lowell Observatory, 1400 West Mars Hill Road, Flagstaff, Arizona 86001, USA
e-mail: rlm@lowell.edu

It should be possible to replace animals in research

SIR — In his Correspondence 'Replacement of animals in research will never be possible' (*Nature* **457,** 147; 2009), Roberto Caminiti makes a case for retaining the current breadth of medical research in using non-human primates. Although immense progress has been made from scientifically well-founded work on non-human primates, I cannot agree with his contention that it will never be possible to replace these animals in research.

To my mind, there is a moral inconsistency attached to studies of higher brain function in non-human primates: namely, the stronger the evidence that non-human primates provide excellent experimental models of human cognition, the stronger the moral case against using them for invasive medical experiments. From this perspective, 'replacement' should be embraced as a future goal.

We should not assume that good medical science is by definition morally justifiable or morally acceptable. The European Union proposal that sparked Caminiti's Correspondence is rekindling this morally and scientifically essential debate.

Bill Crum Centre for Neuroimaging Sciences, Institute of Psychiatry, King's College London, De Crespigny Park, London SE5 8AF, UK
e-mail: bill.crum@iop.kcl.ac.uk

Guarding Hubble telescope's future for posterity

SIR — In her Review '18 years of science with the Hubble Space Telescope' (*Nature* **457,** 41–50, 2009), Julianne Dalcanton discusses the telescope's remarkable achievements. I am saddened at the thought that the best end for Hubble that NASA can devise is simply to burn it up.

Although the coming repair mission may be the last from NASA, it need not be the last mission to Hubble, which will still have use as a scientific instrument. Perhaps another nation might want to adopt it? As humankind progresses farther into space, a saved Hubble would be a treasured artefact.

What is necessary right now is that NASA should use this last mission to secure attachment points to Hubble so that a future, unmanned satellite could dock and raise it to a secure orbit.

Paul L. Schwartz, East Hampton, New York 11937, USA
e-mail: plschwartz@hotmail.com

Benefits of stemming bovine TB need to be demonstrated

SIR — In their replies to our Correspondence on bovine tuberculosis (TB) 'Does risk to humans justify high cost of fighting bovine TB?' (*Nature* **455,** 1029; 2008), Noel Smith and Richard Clifton-Hadley ('Bovine TB: don't get rid of the cat because the mice have gone') and Stephen Gordon ('Bovine TB: stopping disease control would block all live exports') argue that these costs are indeed warranted (*Nature* **456,** 700; 2008).

The current surveillance system for bovine TB may remove cattle at an early stage of infection, as Smith and Clifton-Hadley suggest, but the UK cattle herd is generally young and bovine TB is a chronic disease. Hence, in the absence of control, very few cattle would be likely to reach the advanced stage of the disease at which airborne transmission might, at least in theory, be increased. It is also worth noting that in the early twentieth century, before bovine-TB control was instigated, almost all cases of zoonotic TB in the United Kingdom were non-pulmonary, and most of the small number of pulmonary cases seemed to come from haematogenous spread rather than airborne infections (A. S. Griffith *Tubercle* **18,** 528–543; 1937). With larger farming enterprises and fewer farmers, the at-risk population would be a fraction of that prevailing previously. Human exposure would therefore be likely to remain very low, even in the absence of bovine-TB control.

Gordon claims that abandoning bovine-TB control would bring all live exports to a stop, and he compares Britain's bovine-TB programme to that put in place for foot-and-mouth control in 2001. But foot-and-mouth disease has a devastating economic effect on livestock that goes beyond merely closing down exports, and the cost of the 2001 UK outbreak was a one-off, non-recurring charge. Bovine-TB control is costing up to £99 million (US$140 million) a year and, according to Gordon's own data, live exports have a value that is considerably less.

In addition, the ongoing bovine-TB programme in the United Kingdom is failing to eliminate the disease. Those who propose to spend large amounts of public money on bovine-TB control need to demonstrate the economic and/or public-health benefit; so far, such evidence has been lacking.

Paul Torgerson Ross University School of Veterinary Medicine, PO Box 334, Basseterre, St Kitts, West Indies
e-mail: ptorgerson@rossvet.edu.kn
David Torgerson Department of Health Sciences, University of York, York YO10 5DD, UK

Contributions may be submitted to correspondence@nature.com.

COMMENTARY

Not honouring the code

Countries are not complying with the UN Code of Conduct for Responsible Fisheries. It's time some changes were made, say **Tony Pitcher**, **Daniela Kalikoski**, **Ganapathiraju Pramod** and **Katherine Short**.

A widely agreed remedy for overfishing, which has dramatically depleted fish populations in the world's oceans, would be to adopt the voluntary Code of Conduct for Responsible Fisheries, developed by the Food and Agriculture Organization of the United Nations in 1995[1].

The code provides a detailed consensus for the scientific, sustainable, responsible and equitable exploitation of fishery resources. Now, 13 years after its publication, a detailed evaluation for the 53 countries landing 96% of the global marine catch (based on reported catch in 1999) reveals dismayingly poor compliance. To improve matters, we suggest establishing mandatory instruments, either national or international, that echo the specific requirements for compliance with the code, and tailoring aid for developing countries to address specific weaknesses.

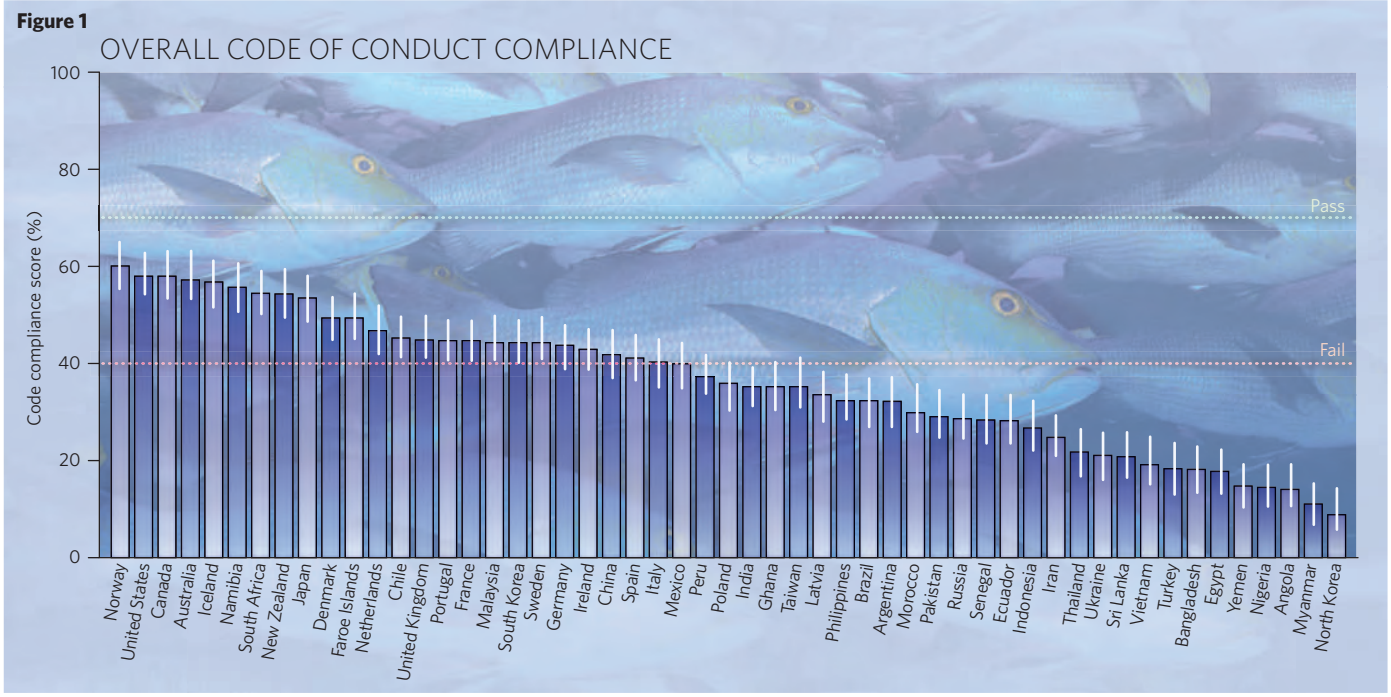
In 2004, we began an extensive analysis of the most active fishing countries in the world. We evaluated the published and unpublished literature, and probed expert opinion to answer 44 questions[2] about adherence to Article 7 of the code, which covers fisheries management, for the 53 countries[3]. The questions fall into six evaluation fields. The first three measure intentions to comply with the code, rating a country's balance of conservation and economic aims; its stated management targets; and its use of precaution when expanding fisheries and establishing no-take zones. The remaining questions deal with the effectiveness of day-to-day compliance, including the rigorous use of quantitative reference points, minimizing wasteful discard, by-catch and impact on habitats such as coral reefs; socio-economic factors such as maintaining beneficial small-scale fisheries and coastal communities; and the control of illegal fishing and 'flags of convenience', when ships are registered in countries other than those where they are owned in order to evade regulation.

"The time has come for a new international legal instrument."

Questions were scored against criteria on a scale of zero to ten, with upper and lower estimates of confidence provided. We considered a score of seven or better to be 'good'; below four a 'fail' grade; and everything in between to be a 'pass'. Although such simple grades can seem arbitrarily chosen, the threshold criteria used for each question were as objective as possible. Scores were cross-checked, subjected to external validation protocols including preliminary publication online as 'living documents'[3], and statistical modelling to deal with uncertainty, all of which is detailed in our final report[4].

Overall, compliance is poor, with room for improvement at every level in the rankings (Fig. 1). Not one country achieves a score in the good category and the average of all countries' ratings barely exceeds the fail threshold. Only six countries have overall compliance scores whose confidence limits overlap with 60% (Norway, the United States, Canada, Australia, Iceland and Namibia), yet four of these top-ranking countries falter by being awarded at least three fail grades, revealing that there is room for improvement even for countries at the top of the rankings. Overall, the five questions on which countries scored worst concerned introducing ecosystem-based management, controlling illegal fishing, reducing excess fishing capacity and minimizing by-catch and destructive fishing practices. At the lower end, 28 countries, representing more than 40% of the world fish catch, had unequivocal fail grades overall. Including confidence limits that overlap the fail threshold raises this to 34 failing countries taking about 60% of the global catch. Twelve countries were awarded fail grades

Figure 1



in all or most parts of the compliance analysis.

Although intent to comply with the code is high in many countries, intentions unsurprisingly exceed compliance by 9% on average (11% in the top-ranking countries). Taking averages for regions of the world, North America (Canada and the United States, n=2 countries) scored towards the top of the pass range in intentions, and in the mid range for implementation. Australasia (n=2) has quite high compliance ratings, with intentions achieving the 'good' range. The averages for African (n=8), Asian (n=20) and Latin American (n=6) countries by contrast, fail in nearly all categories. Although Europe had some of the highest scores, disappointing scores from some European Union nations, with the undoubted resources and know-how to implement the code, reinforce a low priority given to improving fisheries management.

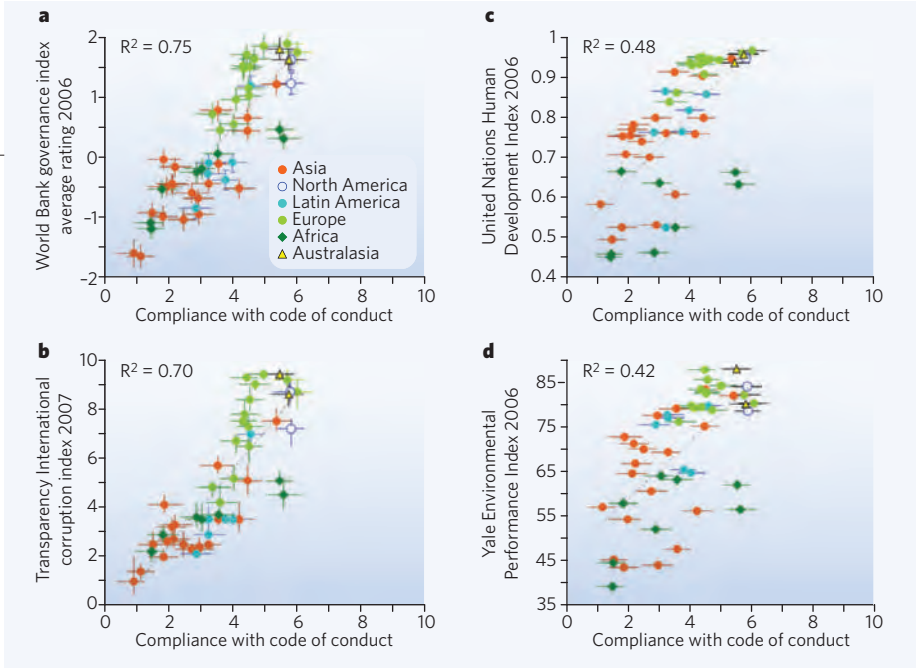


Figure 2 | Tracking trends. Comparing code compliance with **a**, the World Bank Governance Indicators (2006, n=53); **b**, corruption index from Transparency International (2007, n=52); **c**, UN Human Development Index (2006, n=51); **d**, Environmental Performance Index (2006, n=52). Bars indicate confidence limits where available (broken line shows linear trend and R^2, coefficient of determination).

Transparent correlation

Some might argue that we have been generous in awarding scores for published legislation or policy documents intending to comply with the code, and that only actual compliance results should be used to rate countries. The code itself, however, encourages formal legislation and makes this distinction. Moreover, the costs of enforcement can be prohibitive, even in relatively prosperous countries. In Canada for example, government auditors have criticized failure to implement ocean-management legislation: an ironic twist considering Canada pioneered drafting the code in the 1990s.

We have compared code compliance with other relevant indicators of country wealth, governance and environmental performance. There is a strong correlation of estimated code compliance with country scores on the World Bank governance index (Fig. 2), which measures such parameters as political stability, violence, corruption and accountability[5]. Figure 2 also shows similar, if weaker, correlation with Transparency International's Corruption Perceptions Index[6], with the United Nations Human Development Index[7] and, at a lower level, with the Yale Environmental Performance Index[8]. Interestingly, there was no significant correlation between code compliance and size of catch or with the Gini coefficient, an index of socio-economic equity. Both had been thought to relate to poor fisheries management.

Some outliers are worth noting. It is encouraging that some developing countries (Malaysia, South Africa and Namibia, for example) scored more highly than many developed European countries and also more highly than the overall trend, signifying that some elements of good fishery management can be achieved with limited resources. There are considerable differences between these countries, although in each case targeted development aid is likely to have been a factor. Namibia, for example, received fisheries aid from Scandinavia and inherited fairly well-manged fisheries from South Africa.

The poorer compliance of many EU countries than their governance and resources would indicate is possibly partly as a result of a dysfunctional Common Fisheries Policy. Norway and Iceland (both non-EU countries), have much better compliance, perhaps because of heavier reliance on fisheries in their national economies and a long Scandinavian tradition of support for code development — including overseas aid and national measures such as a ban on discarding unwanted fish at sea, effective control of illegal fishing and precautionary fishing targets. The reasons underlying these outliers would make good topics for further research.

We draw two main conclusions from our work. First, compliance scores from developed nations are on average twice as high as those from developing nations, although some notable developing countries with limited resources have scored quite well. Compliance in poorer countries could be boosted through development aid targeted on issues where code compliance is weak. For example, surveillance has been improved in Thai, Moroccan and Malaysian fisheries through aid providing new patrol vessels and modern electronic monitoring devices. Training in quantitative stock assessment and formal management-strategy evaluation using target and limit reference points has been sponsored in several Asian countries.

Second, although the voluntary nature of the code may have been necessary in getting all-nation agreement when it was drafted in the early 1990s, attitudes to the oceans have changed. There is now widespread scientific consensus on the ecological impacts of continued overfishing and the threats to seafood security, and broad agreement on policy issues such as curtailing illegal catches[9] and minimizing the impacts of fishing on marine ecosystems. The time has come for a new integrated international legal instrument covering all aspects of fisheries management. ■

Tony Pitcher and **Ganapathiraju Pramod** are at the Fisheries Centre, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada; **Daniela Kalikoski** is at the Federal University of Rio Grande, Caixa Postal 474, Rio Grande RS, Brazil. **Katherine Short** is at WWF International, CH-1196 Gland, Switzerland.
e-mail: pitcher.t@gmail.com

1. *Code of Conduct for Responsible Fisheries.* (FAO, 1995). Available at: ftp://ftp.fao.org/docrep/fao/005/v9878e/v9878e00.pdf
2. Pitcher, T. J. *Rapfish, A Rapid Appraisal Technique for Fisheries, and its Application to the Code of Conduct for Responsible Fisheries.* FAO Fisheries Circular No. 947 (1999).
3. Pitcher, T. J., Kalikoski, D. & Pramod, G. (eds) *Evaluations of Compliance with the UN Code of Conduct for Responsible Fisheries.* Fisheries Centre Research Reports **14,** 1192 pp. (2006).
4. Pitcher, T. J., Kalikoski, D., Pramod, G. & Short, K. *Safe Conduct? Twelve Years Fishing Under the UN Code.* (WWF, 2009). Available at: http://assets.panda.org/downloads/un_code.pdf
5. www.worldbank.org/wbi/governance/govdata
6. www.transparency.org/cpi
7. http://hdr.undp.org/en/statistics/indices/hdi
8. Esty, D. C. *et al. Pilot 2006 Environmental Performance Index.* (Yale Center for Environmental Law & Policy, 2006). Available at http://www.yale.edu/epi/2006EPI_MainReport.pdf
9. Agnew, D. *et al. PLoS ONE* (in the press).

For full report and to discuss this Commentary, see http://tinyurl.com/aeae6e.

ESSAY

Engineering: Worldwide ebb

In the last in our series on being human, **Melanie Moses** gets to grips with humanity's greatest challenge: how to reduce the demand for energy in increasingly complex, networked and energy-dependent societies.

Humans consume resources equivalent to more than half the production achieved by all the plants and other primary producers on Earth. Our ability to do so, and to distribute those resources across the globe on a scale unparalleled in non-human systems, stems in part from infrastructure networks that connect us to each other and to our environment.

Engineered distribution networks, such as electric-power grids, oil pipelines, railroads, airports, trade routes and banking systems, and the communication networks that their coordination requires, provide the channels through which people, diseases, resources and ideas now move through the world. They determine who we meet, where we travel and how much we consume.

Many of our most pressing global challenges stem from flows over these networks. Air, rail and road networks vastly increase the likelihood of a pandemic by allowing billions to be infected by a virus at an unprecedented speed. Shipping networks carry energy from a few oil-rich locations to distant consumers, fuelling potentially catastrophic climate change. Even less tangible networks have tremendous impact on humanity: most economists failed to predict the speed and extent of the recent financial crisis partly because they didn't understand the nature of the networks through which it spread.

To manage our impact on the environment and understand the ramifications of our actions in an increasingly interconnected world, we need a macroscopic view as well as a detailed understanding of the structure of the networks we have created. The bigger picture is beginning to emerge from theoretical approaches that reveal the structure and dynamics of networks, how networks change as they grow, and how networks constrain individual behaviour.

A question of size

In the past decade, complex network theory has begun to describe the structural features of networks. It shows, for example, that in most social networks, many people have only a few connections whereas a small number of individuals are highly connected. The connections are also clustered: for example, we tend to know lots of our friends' and colleagues' acquaintances, but are less likely to know people from different professions or social groups. Understanding the structure of networks should certainly help reduce the spread of disease, or facilitate the spread of trust in a financial system. But it doesn't shed much light on how physical resources move through networks.

The Metabolic Theory of Ecology (MTE) offers one way to understand the dynamics of flow through networks. The mathematical foundation of MTE was developed a decade ago by a group of biologists and physicists who wanted to explain why so many characteristics of plants and animals systematically depend on their mass in a very peculiar way. The theory posits that much of the life history of an animal (such as how long it lives, how often it reproduces and how much it eats) is determined by geometric and dynamic properties of the cardiovascular network that controls its metabolism.

According to the theory, the larger the animal, the longer its cardiovascular system (its network of arteries and capillaries) takes to deliver resources to its cells. That delivery time, which in turn dictates the animal's metabolic rate, is proportional to the animal's mass raised to the power of ¼. Thus, because its circulatory system works less efficiently, an elephant grows systematically more slowly than a mouse, with a slower heart rate, a lower reproductive rate and a longer lifespan.

> "As infrastructure grows we get more out of it, but must invest more into it, reducing the energy and capital left to invest in the next generation."

Biologists disagree over exactly how much slower the metabolism of larger animals is, and whether the networks-based explanation for the relationship between mass and metabolism is the correct one. However, the implications of this basic idea — that networks become predictably less efficient as they grow — are profound. Indeed, MTE offers insights that could revolutionize the way we understand, predict and manage large networked systems. As well as suggesting that larger systems process energy proportionally more slowly than smaller ones, it implies that the rate at which a system processes energy drives much of its broad-scale behaviour, whether that system is an organism, society or technology.

A common trend

Applying MTE to human social systems sheds light on the well-known but little understood decline in fertility rates that occurs with economic development. As societies consume more energy, people become wealthier but they also have fewer children. Today that energy primarily takes the form of fossil fuels. The average human uses up only about 100 watts from eating food, consistent with predictions based on body size. But in North America, each person uses an additional 10,000 watts from oil, gas, coal and a smattering of renewable sources, all of which are delivered through expansive, expensive infrastructure networks.

The decline in fertility rates with economic growth, called the demographic transition, has puzzled human-life-history theorists for decades: that the people with the most resources have the fewest offspring apparently contradicts basic Darwinian expectations, particularly as the gain in fitness resulting from improved offspring survival is far too small to compensate for the drop in birth rates. But MTE shows that this pattern is not unusual. In fact, across contemporary nations, the decline in human birth rates with increased energy consumption is quantitatively identical to the decline in fertility rate with increased metabolism in other mammals. Put another way, North Americans consume energy at a rate sufficient to sustain a 30,000-kilogram primate, and have offspring at the very slow rate predicted for a beast of this size.

Does the common pattern shared by humans and other mammals point to a common cause? In mammals, the amount invested in each offspring increases with the size of the animal, but, as discussed above, the proportion of energy available for reproduction declines as animals get bigger. Thus, an elephant takes longer (by a factor of mass raised to the power of ¼), than a mouse to acquire the resources needed to reproduce. It may be that humans similarly invest a constant fraction of available resources into each child, but take longer to acquire that fraction in wealthier societies. According to this hypothesis, as our infrastructure grows, we get more out of it, but we must invest proportionally more into it, reducing the energy and capital left to invest in the next generation.

ILLUSTRATION BY G. BECKER

Alternative explanations for the demographic transition include cultural traditions that enable women to have fewer children than they are biologically capable of having, such as birth control or marrying late. But these mechanisms don't explain why women choose to delay having children in the first place. Another theory, based on 'embodied capital', proposes that as societies become wealthier, greater educational investments are made in each child to make them competitive in labour markets. This idea is broadly consistent with the hypothesis outlined above. Interestingly, years spent in education are strongly inversely correlated with fertility rates across nations.

In the face of climate change, the correlation between the decline in birth rates in wealthy nations and a sharp rise in energy consumption is alarming. Even the government-mandated fertility-rate reduction in China, which has resulted in a 70% drop in birth rate in three decades, has been accompanied by a dramatic increase in per capita energy consumption. If such correlations continue to hold, the projected 9 billion people alive in 2050 would achieve zero population growth only if each person consumed close to what the average European consumes today. That would increase total human consumption by ten times.

The recent economic crash demonstrates the fragility of financial networks and the tremendous impact their collapse can have; transportation networks facilitate the spread of disease; and distribution networks deliver sufficient energy to enable us to alter the planet's climate. Yet technological developments continue to connect more of the world's people. Moreover, achieving further economic growth, particularly in developing nations, will probably require more energy and therefore even more expansive networks.

Efficiency drive

Several crucial messages are emerging from early work on human-engineered networks. Human societies are complex systems that persist by consuming energy, but energy consumption cannot be lessened simply by reducing individual demand. Any one person's consumption and, possibly, fertility rate, is affected by structures at higher levels. Relating the behaviour of individuals to global-scale problems will require understanding those individuals as nodes in a network, in which the behaviour of one affects the whole society and where the collective behaviour of the society constrains the behaviour of individuals.

Another key message is that centralized networks, in which resources flow from a central place out to scattered destinations, deliver energy less efficiently when they transport goods over longer distances. Global agricultural production increased six-fold from 1900 to 2000 by increasing energetic investment in agriculture 80-fold. This appalling return on investment in the energy used to fertilize, harvest and transport food means that we now put more energy into acquiring food than we obtain from eating it. Better network designs are critical to ensure that the transportation networks, technologies and economies of the future give far better returns on energetic investment, even as they grow to serve more people. The world's remaining oil reserves are clustered in a few locations far from most consumers. But a more efficient, decentralized infrastructure could be built to deliver energy harvested from wind, solar and tidal resources.

Already, potentially more efficient ways to design infrastructure are emerging. My colleagues and I recently showed that some at least partially decentralized networks, such as computer networks and urban roads in cities (where half the world's population now resides), can increase in size more efficiently than purely centralized ones. For example, our models show that traffic, and so oil consumption, can be proportionally reduced as cities expand, as long as multiple recreational and commercial centres are placed near residential areas. Moreover, Luís Bettencourt and his colleagues recently showed that certain factors, such as innovation and wealth creation, increase superlinearly with city population. In this instance, the more people in a city, the more each person benefits from the collective ability to interact.

In the decades ahead, we need to understand how social and infrastructure networks constrain individual behaviour, and structure cities and societies in ways that increase innovation-inducing interactions but reduce transport and travel distances. By doing so, we'll stand a better chance of meeting the needs of a large, voracious and growing human population without decimating the resources available to future generations. ■

Melanie Moses is at the Department of Computer Science, University of New Mexico, Albuquerque, NM 87131.
e-mail: melaniem@unm.edu

See http://tinyurl.com/d6ck5c for further reading. For more on Being Human, see www.nature.com/nature/focus/beinghuman.

BOOKS & ARTS

Morals and manners in modern science

Today's research enterprise is often portrayed as impersonal and calculating, but a historical examination argues that scientists' civility to each other is what holds the venture together. **Jerome Ravetz** explains.

The Scientific Life: A Moral History of a Late Modern Vocation
by Steven Shapin
University of Chicago Press: 2008.
486 pp. $29

NATIONAL MEDIA MUSEUM/SSPL

Did the early pioneers of industrial science rely on civility and cooperation for success?

As the embodiment of objectivity, scientific knowledge has been placed at the heart of the transition to modern society. Generations of social theorists, including Max Weber, have considered the rise of rationality to be a good thing. But Weber also wrote of his regret of society's loss of 'spirit'. Today, many scientists also harbour a sense of nostalgia for the 'little science' that prevailed before the era of the atomic bomb — science that was small-scale in personnel and resources and happily autonomous. This has been supplanted by the large-scale, capital-intensive science that serves industry and is organized in an industrial mode.

In *The Scientific Life*, historian Steven Shapin asks if contemporary high-tech science is a moral enterprise. Does objectivity render scientific achievement less personal than that in the humanities, and does the scientist possess any special moral virtue? Shapin threads his way through this tangled set of issues with skill, leaving only a few loose ends.

In his 1994 book *A Social History of Truth*, Shapin analysed the importance of social conventions, notably civility, in the constitution of communities that are effective in the pursuit of scientific knowledge. In the seventeenth century, for example, Robert Boyle established the debates of the fledgling Royal Society as a way of securing trust in reports of experimental facts; this, Shapin argues, was the foundation of modern science. *The Scientific Life* moves on his theme of civility to 'late modernity' in connection with contemporary science.

> "In rapidly moving fields such as synthetic biology, scientists must rely heavily on each others' virtue."

Shapin targets the past generation of sociologists of science, including founding father Robert K. Merton. Merton and his followers proclaimed that academic 'pure' science was the only real sort and that those scientists who were involved in the cash nexus were morally inferior. This focus on the purity of science has never been seriously challenged. Commercial scientists are still stigmatized and academic snobbery still rules. Shapin sets out to redress that injustice, with detailed studies of the precepts and practice of both academic and industrial research in the modern world. This change of focus is important because the pockets of science that are unaffected by commerce, or by the state, are steadily shrinking. To ignore this would be to deny the realities of today's science. After this book, that cannot happen again.

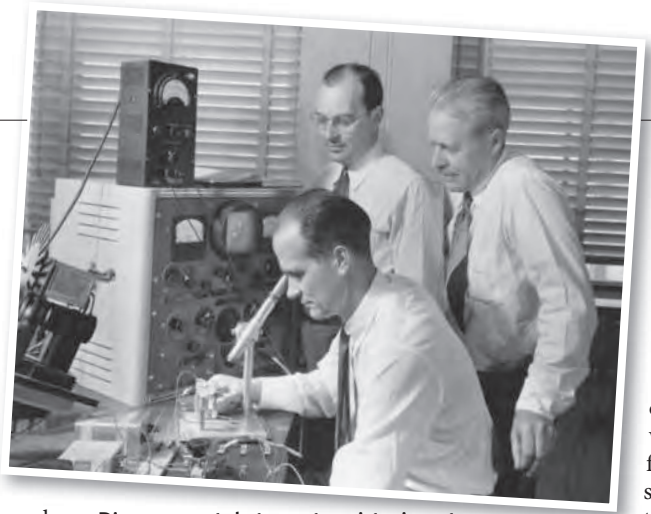
The Scientific Life should therefore be required reading for all scientists and those studying the social activity of science. After expressing the ambiguities and tensions in the scientific role, Shapin shatters myths by contrasting two views of industrial scientists, one from academia and the other from inside industry. It emerges that the founders of the great industrial labs recognized the need for independence and creativity among their scientific workers, providing them with incentives that supported the long-term welfare of the lab. Shapin points out that integrity is appreciated even in industrial science, before going on to focus on the role of morality in teamwork and in the planning of research. Because of the radical uncertainty of the 'future-making practices' of speculative, commercially oriented science, Shapin argues that the virtues of the people involved are all that one has to rely on in setting a path for the advancement of research.

Shapin sums up his argument with an anecdote: describing a farewell party at the University of California in San Diego, he is impressed by the crowd's civility. Differences in rank, prestige, wealth and influence are ignored; tact and consideration rule. It is the personal element that governs this most modern of enterprises.

Shapin uses powerful terms: moral, virtue, vocation, charisma. But he focuses on the positive attributes of social decencies. He neglects that in an individual these attributes are achieved by struggle and sacrifice, and in a group they also refer to its collective activities. Although he provides full and illuminating accounts of the social practices of different areas in contemporary science, Shapin fails to distinguish between politeness and civility on the one hand and morality on the other. And by this omission he presents an invitation to his critics.

The dark side

As a historian, Shapin will know of counter-examples to his thesis that 'manners maketh man'. Most notable is the impression created by the Royal Society on the young writer Voltaire during his visit to England in 1727. Voltaire remarked most favourably on the gravity and courtesy of the English, a civility — in the tradition of Boyle — that was so refreshingly different from the disputatious style of the French. Yet a few insiders at the Royal Society then knew a secret that would wait for more than a century to be exposed: the society's


BELL LABS/AIP EMILIO SEGRE VISUAL ARCHIVES

Disagreements between transistor inventors William Shockley (seated), John Bardeen (left) and Walter Brattain ended their fruitful collaboration.

good name had recently been used in the character assassination of Gottfried Leibniz, the German philosophical genius and protégé of the future King George I of Great Britain and Ireland. Isaac Newton himself had secretly masterminded the attack, personally embellishing the society's dossier about Leibniz's supposed plagiarism of the calculus during a visit to London in 1676. In this case, civility did not ensure morality, of any sort.

Shapin is completely aware of this dark side of science. His excellent bibliography lists the writings of many respected voices, such as Sheldon Krimsky in *Science in the Private Interest*, complaining about morally dubious practices in science. But their critical perspective never appears in his narrative. One would not know from reading *The Scientific Life* that Shapin has published many essays showing deep sympathy with those who offer such complaint. For example, he writes of Craig Venter, who has pioneered the creation of new life forms for private profit. He answers objections as they are raised, but the deeper issues of the safety and morality of Venter's enterprise are ignored.

Shapin's study is neither a sources-based history of the past nor an empirical social-science analysis of the present. It is instead an extended insightful essay. This genre enriches public debate — be it by an academic, as in David Riesman's *The Lonely Crowd* or Robert Putnam's *Bowling Alone*, or the product of a distinguished journalist, such as James Fallows's classic *National Defense*. Through his many writings on science, Shapin has become one of these public intellectuals. But Shapin's book lacks a critical edge. It is as if he has been so seduced by civility, ancient and modern, that he has devoted his great talents to extolling its virtues.

For a scholar of Shapin's stature, it is inappropriate simply to say that he has forgotten his own critical awareness. He is serious in his use of ethically charged terms. I see this controversial element as a symptom of an unresolved problem in a bigger endeavour: those who promoted the organization of societies around objective scientific rationality, such as Weber, were not talking merely about more education and more knowledge. In the Enlightenment movement that flourished through the eighteenth and nineteenth centuries, science was seen as the main weapon against the obscurantism that drew on dogma and superstition. Impersonal science was therefore seen as the key to both real knowledge and a good society. In the intervening period, that early vision of science has been badly damaged. Today, public distrust has become widespread.

Shapin, however, believes that in today's industrialized science, the politeness and civility of pre-modern communities persist. They are essential, he thinks, because of the uncertainties that beset this science. In rapidly moving fields such as synthetic biology, scientists must rely heavily on each others' virtue. In that way, for him, even scientific entrepreneurs are creating microcosms of a good society. Neither modernity nor industrialization needs to be dehumanizing. Regardless of the imperfections of his current evidence, and the counter-examples that can be adduced, that is a thesis that deserves respect and critical engagement.

I have a final reflection. Had Shapin chosen to study the mathematicians who are employed in the world of finance, he might well have found similar patterns of civilized interaction and similar evidence of individual moral virtues. Yet we now know that the collective endeavour of these other very nice entrepreneurial scientists has resulted in the creation of a mountain of toxic fake securities. A sobering thought. ■

Jerome Ravetz is associate fellow at the James Martin Institute for Science and Civilization at the University of Oxford, Oxford OX1 1HP, UK. He is author of *Scientific Knowledge and Its Social Problems*.
e-mail: jerome-ravetz@tiscali.co.uk

Natural selection and the nation

Banquet at Delmonico's: Great Minds, the Gilded Age, and the Triumph of Evolution in America
by Barry Werth
Random House: 2009. 400 pp. $27

"There is apparently much truth in the belief that the wonderful progress of the United States, as well as the character of the people, are the results of natural selection," wrote Charles Darwin in *The Descent of Man*. Today, such a claim jars, and not just because of the grammar. But the wonder is that Darwin so infrequently over-extended his evolutionary explanations.

Not so his English contemporary, Herbert Spencer, who attempted to give an evolutionary account of almost every realm of human affairs. Ruminating on history, psychology, sociology and ethics, Spencer's evolutionary philosophy led him to argue that, among other things, government regulation was bad, the poor and needy should be left to fend for themselves, and the United States was destined to become the pinnacle of civilization. These ideas fell on fertile ground, particularly in the United States, and Spencer was hailed there as the brightest, most insightful man of his generation.

Banquet at Delmonico's, titled after an 1882 dinner to honour Spencer at a New York restaurant, covers the elite's battle for ideas during the turbulent years of the 1870s and 1880s. The nation was emerging from a bitter civil war that had led many to question the benevolence of God. It was obvious that the country was about to transform itself from an underpopulated minor player to a world-dominating industrial giant, and its direction and politics were up for grabs. The issues of the times were challenging: credit crunches, presidential unpopularity, disputed elections, terrorist atrocities, military blunders, and arguments about the nature of marriage, race relations and intelligent design. Manhattan got electrical lighting, Pittsburgh got steel and General Custer got annihilated.

Spencer's US acolytes included powerful industrialists, politicians, religious leaders and intellectuals. In a beautifully written classic of non-fiction narrative, author Barry Werth tracks Spencer and associated characters as they try to use evolutionary doctrine to perfect humankind and society, often attempting to take the credit. The startling cast includes the liberal Christian minister and alleged adulterer Henry Ward Beecher, the first female candidate for US president, Victoria Woodhull, and the publisher and self-flagellating scientific crusader Edward Youmans. Among the academics are Harvard University's John Fiske, who believed in the country's divine destiny, Louis Agassiz, who believed human races were

Herbert Spencer felt evolution could cure social ills.

J. BAGNOLD BURGESS/NATIONAL PORTRAIT GALLERY, LONDON

created separately, and Asa Gray, who believed fervently in both Darwin and Christianity. The book is so rich in details — church meetings, fossil hunts, ocean voyages, hikes, courtroom dramas and Victorian hypocrisy — that it reads like a novel. But the narrative drive is weak: it is often hard to see where the story is going or why. That, I guess, is reality.

It is also not obvious why the book culminates with much hyperbole in the eponymous banquet. This 12-course meal, with a separate wine for each course, was held at a famous Manhattan restaurant shortly after Darwin's death. It was attended by 200 of the most powerful men in the United States, and celebrated Spencer at the end of what was to be his last US trip. The build-up to the meal is tremendous, and we are treated to all the procedural details — course three of the first service included buttery, scarlet kettle-drum-shaped pastry tufts stuffed with truffles, tongue and pistachios — and there is a very full summary of the three hours of after-dinner speeches. The book's cover claims that this event was "a historic celebration from which the repercussions still ripple throughout our society". But I now understand why I had never heard of it. Spencer himself found the speeches boring and wanted to leave early. The audience found a new idea only in John Fiske's speech: he asserted that humans acquired a sense of morality not from God, but from natural selection. The only speech that might resonate today was Spencer's own. Worried about the country's well-being and health, he railed against the national work ethic, arguing that Americans should spend less time striving for a future good, and more time enjoying what the passing day had to offer. The idea baffled his audience and was poorly received.

Yet the narrative non-fiction genre allows unexpected things to emerge. Many of the protagonists were, like Darwin, bedevilled by bad health. Doctors are summoned at a frightening rate throughout the book. The ailments were many and the treatments fascinating — at one point, Agassiz was forbidden from thinking. That natural selection should have excised such sick men does not seem to have caused much concern among any of these social Darwinists. Moreover, neither Spencer nor any of his US disciples seems to have spent any time trying to push evolution into medicine, even though medicine was becoming a serious scientific enterprise, with the verification of the germ theory of disease and the developing cellular theory of disease (now pathology). Even today, medicine is the most obvious area in which evolutionary biology remains under-extended. Mutation, competition and selection are key to understanding cancer and infectious diseases, for example, but still very few medical schools teach evolutionary biology.

We have yet to fully comprehend the consequences of what Darwin did to humanity's view of itself. Werth's picture of what his 'great minds of the gilded age' were thinking, of how far they tried to stretch Darwinian insights, and of the personal and moral lessons they drew, makes a forceful argument that the causes of biological diversity — and humankind's place within it — really matter. The fact that many of these thinkers' conclusions were based on such a poor understanding of evolution also shows why everyone deserves proper schooling in evolutionary biology. The Victorians had the crippling disadvantage that they did not understand inheritance or units of selection. Today, humanity has no such excuse. ■

Andrew F. Read is professor of biology and entomology at Pennsylvania State University, University Park, Pennsylvania 16802, USA.
e-mail: a.read@psu.edu

Portraying the embryo

Making Visible Embryos
by Tatjana Buklijas and Nick Hopwood
Exhibition at http://tinyurl.com/9m8s7u

"Do we not find a rosebud as beautiful in its own way as a rose?", mused the great German anatomist Samuel Thomas Sömmerring. He was defending his revolutionary work, published as a series of large-format plates in 1799, showing that embryos took different forms at different stages.

The story of how embryos have been depicted is the subject of the online exhibition *Making Visible Embryos*, by historians of science Tatjana Buklijas and Nick Hopwood.

Before Sömmerring, anatomists adhered to the Aristotelian theory that the individual adult was present in the germ cell, and simply grew in size — no rosebuds, just small, perfect roses. The debate was only whether the homunculus was encapsulated in the egg or the sperm. The concept of the embryo as an unformed blob did not fit with theories of the Creator's perfection.

Then experimentation took over. Human embryos were in short supply, but Sömmerring systematically acquired them from abortions, picked out the best examples, which he assumed to be less malformed, and drew his own conclusions.

The exhibition of more than 120 images, from engravings and wax models to X-rays and ultrasound scans, presents how scientists have struggled to understand the embryo in its biological and moral contexts. We learn how Jesus was often depicted as a small but fully formed child in the womb of the Virgin in medieval and early-modern paintings. We learn how German experimental zoologist Ernst Haeckel, one of Charles Darwin's most insistent propagandists, used his considerable artistic skills to present images of developing embryos — and massaged some to support his theory that different species pass though similar embryonic stages. And we discover how the emotionally powerful images of Swedish photographer Lennart Nilsson were, ironically, taken from aborted fetuses. In the 1960s, these photographs influenced the modern public view of the fetus as a child waiting to be loved, and thus fuelled the fire of anti-abortionists.

The website is structured by theme; each section runs chronologically and information is provided at three levels of depth. This architecture mostly works well. But it is less suited to complex discussions, such as the nineteenth-century scientific controversies over embryology, in which it is easy to lose track of the different players, their arguments and how it all fitted together. But the pictures speak volumes, even though images of the embryo are nowadays commonplace. ■

Scientists initially struggled to grasp how embryos develop.

ANATOMISCHES INSTITUT, BASEL

Alison Abbott is *Nature*'s senior European correspondent.

More than skin deep

From patriotism to martyrdom to surgery, Andrew Krasnow's American flag made from human skin reveals many layers of what it is to be human, finds **Martin Kemp**.

Flags are not benign human inventions. At their best they identify communities that give support for meaningful human lives. But at their frequent worst they are waved aggressively in the faces of outsiders. The schematic heraldry of flags is both primitive and enduringly potent.

At one level, flags are robustly unambiguous, but in the emotional domain they are open to polarized readings. For patriotic Americans who attended the inauguration of President Barack Obama, the Stars and Stripes connotes what is good about living in one of the 50 states. For those who collectively or individually see the United States in an opposed light, the flag looks 'ugly'.

When an artist takes on the Stars and Stripes, as Jasper Johns did with his famous series of thickly painted flags from 1954 onwards, openness of reading takes over. Was Johns being patriotic or anti-patriotic? Was he treating the flag as a 'pop' object like Andy Warhol's *Brillo Boxes*? The paintings themselves do not make definite statements.

What are we to make of a Stars and Stripes made out of human skin, as artist Andrew Krasnow has created in *Flag from Flag Poll*? Measuring almost 2 metres wide, it cannot fail to provoke extreme reactions. We are used to artist's provocations, and may suspect that they are all too often more concerned with notoriety than sincerity. How can we judge what Krasnow is doing?

Among other productions, he has manufactured a series of familiar items from human skin — cowboy boots, gun shells and a map of the United States. In his skin hamburger, the bun is equipped with its own set of upper and lower teeth; beef bites back, it seems. He is clearly working around the theme of American icons. We are implicitly invited to take a political stance, and we sense that it is more critical than laudatory. We are hardly likely to read his flag as a militaristic celebration of the American dead but, on second thoughts, such a reading might raise productive ironies.

Because he is using human skin, Krasnow must adhere to legal requirements, but he insists that his supplies were legally obtained 20 years ago. By mounting an exhibition containing "relevant material" that "consists of or includes human cells", GV Art of London, the current home of the flesh flag, has to comply with the requirements of the UK Human Tissue Authority. Set up under the Human Tissue Act 2004, the authority regulates the removal, storage, use and disposal of human bodies, organs and tissues from the living and deceased. In the United States, Krasnow is also potentially violating

GV ART, LONDON

The flesh flag: sensationalist or sincere?

Section 1 of Title 4 of the United States Code, which prohibits desecration of the American flag, even though violation carries no stated penalty and comes into obvious conflict with the right to freedom of speech.

Whatever the legal implications, he tramples on taboos. We may be happy with gloves made from finest kid leather, looking much like the tanned skin used by Krasnow, but surely not if they were made from human skin. A worshipper may treasure a fragment of divine skin from a revered saint placed in a reliquary in a Spanish chapel, but we are unlikely to want such an item decorating our dinner table. We admire wonderful medieval illuminations on vellum ('veal' skin) and ancient leather-bound books, but are disconcerted to find that in the library of Brown University in Providence, Rhode Island, the 1568 edition of Andreas Vesalius's anatomical masterpiece, *De Humani Corporis Fabrica*, is bound in human skin.

Krasnow is working in a medium that is as multivalent in potential meanings and emotional reactions as any art medium could be, even before any content enters. Flaying is full of traditional resonances. In ancient mythology, Marsyas was notoriously skinned for the temerity of his musical challenge to Apollo. St Bartholomew was excruciatingly martyred by flaying. In Michelangelo's *The Last Judgment* in the Sistine Chapel in Vatican City, Italy, the knife-wielding saint, fully intact in heaven, holds his own earthly hide, which bears a distorted imprint of the artist's own face, as Krasnow notes in the exhibition catalogue. Whether or not a Christian is resurrected in his or her own flesh is a long-standing matter of doctrinal dispute.

Closer to home, Krasnow is inviting us to think about "the skin history ... of the Americas, from the scalping of Native Americans on the frontier to the branding of slaves in the south, to the dropping of the A-bomb on Japan to the use of napalm in Vietnam and phosphorus bombs in Central America". He also tells us that his sister "suffered severe burns", from which she died in spite of extensive skin grafts from their parents. However, he resists the idea that there is a dominant "defining experience", either for him or for us.

I am convinced that his art is not merely sensational. He delves into a series of deep and complex strata in our awareness of what it is to be human, individually and under whatever banner we navigate the seas of life. ■

Martin Kemp is emeritus professor in history of art at the University of Oxford, Oxford, UK.

Flag from Flag Poll
by Andrew Krasnow
At GV Art, 49 Chiltern Street, London, until 17 March. See www.gvart.co.uk for details.

NEWS & VIEWS

SOLID-STATE PHYSICS

Electrons in the fast lane

Henning Sirringhaus

Organic semiconductors that operate through the conduction of positive charges are the first choice for use in printable electronic circuitry. A device that uses electrons instead has just joined the rankings.

Transistors, the semiconductor electronic switches at the heart of any integrated circuit, come in two main flavours: p-type transistors, which switch on when a negative voltage is applied to the device's control electrode (gate); and n-type, which switch on with a positive gate voltage. Most silicon integrated circuits make use of a combination of the two types to produce complementary circuits. Such circuits can achieve higher levels of performance and yield, and lower power consumption, than can circuits made from only one type of transistor. The field of organic electronics aims to use as the semiconductor material — instead of inorganic silicon — organic molecules and polymers that can be processed from solution. That would allow large-area electronic circuits to be manufactured on flexible, plastic substrates using low-cost printing techniques[1,2]. Although some p-type organic transistors have the required chemical and physical properties to produce such printable electronic circuitry, researchers have so far failed to find materials for equivalent n-type devices that would offer a comparable level of performance. On page 679 of this issue, Yan *et al.*[3] describe a polymer material that achieves this feat, allowing printed complementary circuits to be produced that have unprecedented performance.

In a semiconductor, whether it is organic or inorganic, two types of charge carrier convey the electrical current: extra, negatively charged electrons that are injected into the empty electronic energy levels of the neutral ground state of the semiconductor, and 'missing' electrons that are removed from one of the normally occupied states. The latter are called holes, and behave as if they were carrying a single positive electron charge. n-Type (n for negative) organic transistors rely on the injection of extra electrons into the lowest unoccupied molecular orbital (LUMO) of the molecules of a thin film of organic semiconductor. Electron injection is achieved by applying a positive voltage to the gate electrode, which is separated from the semiconductor film by a thin gate dielectric (a nonconducting material) (Fig. 1a). p-Type (p for positive) transistors rely on inducing holes in the highest occupied molecular orbital (HOMO) through the application of a negative gate voltage (Fig. 1b).

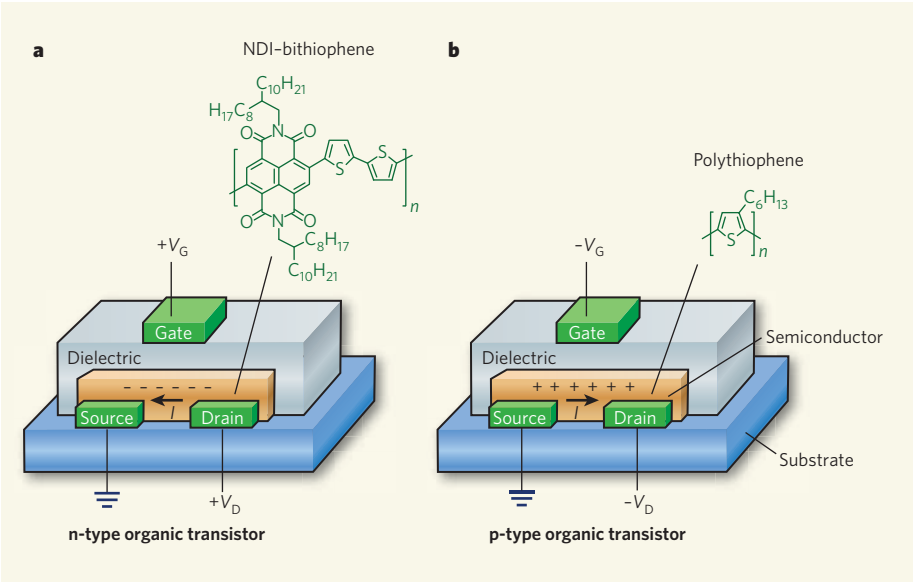


Figure 1 | Structure of organic transistors. The diagrams show the elements that make up the polymer transistors studied by Yan *et al.*[3]. A semiconductor–dielectric layer is deposited on a plastic substrate. **a**, An n-type transistor with a naphthalenecarboxydiimide (NDI)–bithiophene co-polymer as the semiconductor. **b**, A p-type transistor with polythiophene as the semiconductor. The current (I) that flows between the source and drain electrodes in the semiconductor, when a drain voltage (V_D) is applied, is due to electrons (**a**) and holes (**b**) induced by the positive (**a**) and negative (**b**) gate voltages (V_G).

Because the electronic structure of the HOMO and LUMO states is similar, it should, at least in principle, be possible to have both n-type and p-type transistors in a single organic semiconductor. And indeed, a few years ago it was found that many organic semiconductors are inherently capable of both types of device operation when constructed with a suitable gate dielectric[4].

The main challenge to producing technologically useful n-type organic transistors is to ensure adequate operational stability when the device is exposed to the atmosphere. This requires specially designed materials in which the energy of the molecules' LUMO states are stable enough to ensure that the extra electrons do not react with or chemically reduce atmospheric oxygen, water or other electronegative impurities[5] (those that have the ability to attract electrons towards themselves); otherwise, these extra electrons would be lost as current-carrying mobile charges in the device. Organic semiconductors[6] capable of stable n-type device operation are anything but scarce, but most are small organic molecules that are not sufficiently soluble in common organic solvents to be suitable for print-based manufacturing. Polymer semiconductors are useful in this respect, but to date no polymer has shown a stable n-type performance and processability comparable to those of the best p-type polymer transistors.

The new material reported by Yan *et al.*[3] is a donor–acceptor co-polymer. It consists of alternate units of an electron-rich bithiophene donor linked to an electron-deficient naphthalenecarboxydiimide (NDI) acceptor (Fig. 1a). The latter provides the necessary stabilization of the LUMO level. Some of the donor–acceptor polymers reported previously are poorly soluble[7], presumably because strong intermolecular interactions occur between donor and acceptor units on adjacent polymer chains. But the

NATURE|Vol 457|5 February 2009

authors demonstrate that the new polymer is highly soluble in common organic solvents, and is compatible with inkjet, flexographic and gravure printing. Yan and colleagues' transistor has a 'figure of merit' — indicating the mobility of the charge carriers, which determines how much current flows in the semiconductor in response to the applied gate voltage — of 0.4–0.8 $cm^2 V^{-1} s^{-1}$. This value is comparable to that of the best p-type polymer transistors[8].

The NDI-based co-polymer is similar to a previously reported co-polymer that was based on perylenecarboxydiimide and bithiophene, but that was much less efficient[9]. Yan and colleagues' work thus illustrates how subtle variations in chemical structure can have a major influence on the electronic properties of a material. What's more, because the polymer is compatible with the gate dielectrics commonly used for p-type transistors, complementary circuits can be created. The authors put that into practice and demonstrate the feasibility of complementary logic inverters — devices that turn a high input voltage into a low input voltage, and vice versa.

Yan and colleagues' work is a major advance in the quest for printed complementary logic circuits. These can find applications in automatic object-identification techniques, such as electronic barcodes and radio-frequency identification tagging[10,11]. It remains to be established whether the long-term environmental and operational stability of the NDI polymer matches that of the best p-type devices. If it does, the material might also find use in applications such as flexible displays, for which p-type organic transistors are currently the first choice because of their high operational stability. The new material will undoubtedly inspire further work to synthesize other donor–acceptor polymers with even stronger stabilization of the LUMO level.

But the authors' work is not only of interest from a practical perspective. It will also help to refine our understanding of how the physics of electron transport and injection in high-mobility polymer semiconductors relates to that of holes. Electrons in polymer semiconductors are catching up with holes in the fast lane. ■

Henning Sirringhaus is at the Cavendish Laboratory, University of Cambridge, Thomson Avenue, Cambridge CB3 0HE, UK.
e-mail: hs220@cam.ac.uk

1. Klauk, H. *Organic Electronics: Materials, Manufacturing and Applications* (Wiley-VCH, 2006).
2. Crone, B. *et al. Nature* **403,** 521–523 (2000).
3. Yan, H. *et al. Nature* **457,** 679–686 (2009).
4. Chua, L. L. *et al. Nature* **434,** 194–199 (2005).
5. de Leeuw, D. M., Simenon, M. M. J., Brown, A. R. & Einerhand, R. E. F. *Synth. Metals* **87,** 53–58 (1997).
6. Jones, B. A., Facchetti, A., Wasielewski, M. R. & Marks, T. J. *J. Am. Chem. Soc.* **129,** 15259–15278 (2007).
7. Babel, A. & Jenekhe, S. A. *Adv. Mater.* **14,** 371–374 (2002).
8. McCulloch, I. *et al. Nature Mater.* **5,** 328–333 (2006).
9. Huttner, S., Sommer, M. & Thelakkat, M. *Appl. Phys. Lett.* **92,** 093302/1 (2008).
10. Cantatore, E. *et al. IEEE J. Solid State Circuits* **42,** 84–92 (2007).
11. Klauk, H., Zschieschang, U., Pflaum, J. & Halik, M. *Nature* **445,** 745–748 (2007).

CELL BIOLOGY

How to combat stress

Christopher V. Nicchitta

Life is full of stress, and all life forms — from bacteria to humans — have evolved ways of sensing and responding to it. The latest findings shed light on how cells deal with stress.

In cells, protein homeostasis — a delicate balance between maintaining protein conformations, refolding misfolded proteins and degrading damaged proteins — is normally maintained by regulatory networks that control protein synthesis and degradation. Moreover, molecular chaperones are key players in protein homeostasis, helping proteins to fold and preventing aggregation of misfolded proteins, which could have substantial, disease-related consequences[1,2]. So when environmental stress such as nutrient deprivation or oxygen shortage disrupts protein homeostasis, the cell responds. Nowhere is this process more exquisitely controlled than in the endoplasmic reticulum, an extensive organelle consisting of interconnecting tubules that serves as the synthesis site for secretory and membrane proteins. In two fascinating studies from the same group, published in this issue, Korennykh *et al.*[3] (page 687) and Aragón *et al.*[4] (page 736) show that stress elicits the assembly at the endoplasmic reticulum of signalling centres that sense the accumulation of unfolded proteins and direct the appropriate response.

When stress begins to take its toll in a cell and unfolded proteins accumulate, the transmembrane protein Ire1p seems to be the first point of call. The domain of this protein that faces the lumen of the endoplasmic reticulum binds to unfolded proteins there. Its cytoplasmic domain has endonuclease enzymatic activity, and so can cleave the messenger RNA for the vital stress-response protein Hac1p at two specific locations. The result is removal of a domain that prevents Hac1p synthesis, and the rejoining of the two mRNA fragments, to yield an mRNA that can then be efficiently translated[5].

In the first of the two papers, Korennykh *et al.*[3] present a compendium of structural biology and biochemical studies on the Ire1p cytoplasmic domain. What they propose is remarkable: a transmembrane communication event that promotes oligomerization of the cytoplasmic domain into an unusual

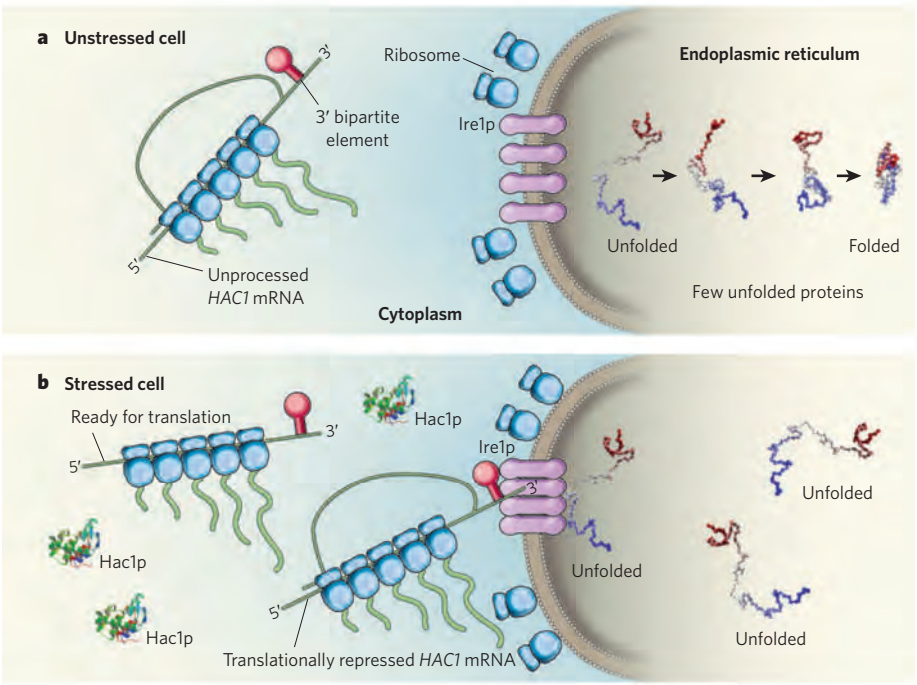


Figure 1 | Emergency response at the endoplasmic reticulum. a, In the unstressed cell, protein folding within the endoplasmic reticulum occurs efficiently; the unfolded protein sensor and signalling protein Ire1p is inactive, and the *HAC1* messenger RNA remains untranslated in the cytoplasm. **b,** Under stress, protein folding is disrupted and unfolded proteins accumulate in the endoplasmic reticulum. It emerges[3,4] that the binding of unfolded proteins to Ire1p promotes clustering of this protein and activation of the endonuclease activity of its cytosolic domain. *HAC1* mRNAs, themselves attached to multiple ribosomes, are then recruited to the activated Ire1p clusters, and are processed there, allowing translation of the essential stress-response factor Hac1p protein.

extended helical rod, much like a DNA double helix. Consequently, the cytoplasmic domain can function as both a kinase and an endonuclease. The authors' structural models are provocative, as they suggest that supramolecular organization is necessary for Ire1p transition to an active, stress-signalling state.

The structure of Ire1p was also reported by another group last year[6]. Although there are clear similarities in the two structures[3,6], the differences are substantial. Unlike Korennykh and colleagues' proposal that higher-order oligomerization is required for activation of the Ire1p cytoplasmic domain, the earlier structure indicated a requirement for a back-to-back dimer arrangement. This inconsistency probably reflects relatively modest differences in the precise domains selected in each study, although Korennykh *et al.* provide mechanistic support for the validity of their model with detailed biochemical studies. In the rest of my article, however, I will focus on the biology of the cellular response to stress-associated accumulation of unfolded proteins.

Aragón *et al.*[4] demonstrate that, on sensing unfolded proteins in the lumen of the endoplasmic reticulum, Ire1p molecules coalesce into large, interacting clusters, and at the same time their cytoplasmic domain becomes active. But captivating questions in the story of the unfolded-protein response are manifold, and the authors provide a surprising answer to one of the more vexing ones: how do Ire1p clusters interact with the *HAC1* mRNA?

The first clue came with the observation that regions in *HAC1* mRNA outside the sites cleaved by Ire1p are necessary for the stress response *in vivo*, but not for mRNA processing *in vitro*. The authors[4] identify one such key region, which they call the 3′ bipartite element (Fig. 1). They find that, in stressed yeast cells, *HAC1* mRNA lacking the 3′ bipartite element interacts with Ire1p clusters only weakly, if at all. Consequently, a rather dismal unfolded-protein response is elicited and the stressed cells cannot sustain growth.

Intriguingly, for the 3′ bipartite element to direct *HAC1* mRNA to the clusters of activated Ire1p, translation of *HAC1* mRNA must be repressed. Aragón *et al.*[4] provide a satisfying answer to why this might be so: with mRNA processing linked to translational repression, only those mRNAs that contain the inhibitory domain find their way to Ire1p clusters. This paradigm provides the first example of mRNA localization serving as a crucial regulatory switch.

As for the remaining questions, perhaps the most compelling is how *HAC1* mRNAs find their way to activated Ire1p signalling clusters. Aragón *et al.* speculate that *HAC1* mRNAs travel from their cytoplasmic location along cytoskeletal filaments to these signalling sites at the endoplasmic reticulum. This is an attractive idea, and regulation of mRNA localization by the cytoskeleton and molecular motors certainly has ample precedent. But would it mean that activated Ire1p clusters also serve as sites for the attachment of cytoskeletal filaments?

And, once on the endoplasmic reticulum, what serves as the binding partner for *HAC1* mRNA? After all, the vast majority of mRNA localization signals are recognized by proteins that contain evolutionarily conserved RNA recognition motifs. So it is the complex of mRNA and RNA-binding protein that directs the localization event. Korennykh *et al.*[3] provide intriguing speculation on this question. They propose that conformational changes in Ire1p that allow its clustering and activation also create a direct binding site for *HAC1* mRNA — potentially another remarkable twist in the biology of mRNA localization.

But if activated Ire1p provides both the binding site for the 3′ bipartite element and the enzymatic function necessary for the *HAC1* mRNA processing upstream of this element, how do unprocessed *HAC1* mRNAs gain access to activated signalling centres? One possibility is that Ire1p-processed *HAC1* mRNAs are poor Ire1p binding partners, and simply diffuse away. Yet the data presented[4] indicate seemingly stable co-localization of *HAC1* mRNAs with the active Ire1p clusters. Perhaps there are other binding partners for the *HAC1* mRNA on the endoplasmic reticulum. Many mRNAs that don't encode secretory or membrane proteins are localized to the endoplasmic reticulum and are translated on ribosomes bound to this organelle[7]. Given this precedent, might the unprocessed *HAC1* mRNAs initially bind to separate sites on the endoplasmic reticulum membrane and be rapidly, and reversibly, transferred to Ire1p signalling clusters? The kinetics of surface chemistry would favour this latter possibility, as reactions occur significantly faster when constrained to a two-dimensional surface rather than in the three dimensions of a solution.

From these lines of questioning, one point should be clear. These ground-breaking studies[3,4] have created both fertile territory for research into the regulation of mRNA localization and mRNA-mediated signalling, and opportunities to gain insights into the fundamental mysteries of the cellular response to environmental stress. The unfolding story of the stress response in the endoplasmic reticulum has yielded many a new paradigm; these latest findings give every indication that the store of surprises is far from exhausted. ■

Christopher V. Nicchitta is in the Department of Cell Biology, Duke University Medical Center, Durham, North Carolina 27710, USA.
e-mail: c.nicchitta@cellbio.duke.edu

1. Morimoto, R. I. *Genes Dev.* **22,** 1427–1438 (2008).
2. Zhang, K. & Kaufman, R. J. *Nature* **454,** 455–462 (2008).
3. Korennykh, A. V. *et al. Nature* **457,** 687–693 (2009).
4. Aragón, T. *et al. Nature* **457,** 736–740 (2009).
5. Chapman, R., Sidrauski, C. & Walter, P. *Annu. Rev. Cell Dev. Biol.* **14,** 459–485 (1998).
6. Lee, K. P. K. *et al. Cell* **132,** 89–100 (2008).
7. Lerner, R. S. *et al. RNA* **9,** 1123–1137 (2003).

CLIMATE CHANGE

Snakes tell a torrid tale

Matthew Huber

The discovery in Colombia of a giant species of fossil snake is news in itself. But a wider, more controversial inference to be drawn is that tropical climate in the past was not buffered from global warming.

As the world uneasily eyes a warmer future, a large community of researchers is investigating the past for the insights it might provide into the likely magnitude of climatic and ecological change. Time intervals in Earth's past, such as the early Palaeogene (between 65 million and 40 million years ago), are known to have been much warmer than today. The presence of fossil crocodiles[1] and palm trees[2] ringing the Arctic and in the hinterlands of Wyoming and Siberia, combined with quantitative records of palaeoclimate, indicate[1–5] above-freezing winter conditions and annual average temperatures in these regions that were often at least 15 °C. But if the extratropics were this warm, how hot were the tropics? Head *et al.* (page 715 of this issue) provide tantalizing clues from an unusual source[6].

Establishing the magnitude of past variation in tropical climate is a formidable challenge. Twenty years ago we thought that the tropics cooled as the world warmed (and vice versa)[7]. Ten years ago the consensus became that, compared with modern values, tropical temperatures were at most only slightly warmer during the various hot, 'greenhouse' climates that have occurred over the past 145 million years[8] and that they cooled by at most a couple of degrees during the ice ages. This muted variation in tropical climate is a puzzle: mechanisms that drive climate change at higher latitudes should also substantially affect lower latitudes. In the early Palaeogene, how could the poles be 30 °C or more warmer than they are today if the tropics were only 2 °C warmer?

For their part, climate modellers have concluded that hot tropical temperatures, and the high concentrations of greenhouse gases that cause them, are required to reproduce warm extratropics, because standard models and

dynamical theory do not produce Equator-to-pole temperature gradients much weaker than they have been in modern times[9]. Nevertheless, on the basis of the supposition that the models are missing important physics, many hypotheses and novel mechanisms have been proposed that centre on the existence of a 'thermostat' that maintains tropical temperatures at a fixed level[10,11]. These attempts to include new feedbacks have illuminated many dark alleys of climate dynamics, but so far all have been dead ends[10].

Whether a tropical thermostat exists is fundamentally important for three reasons. The tropics, defined broadly (30° N to 30° S), make up half of Earth's surface area and so play an outsized part in determining past variations in global mean temperature and the sensitivity of this variable to forcing factors such as greenhouse-gas concentrations. The tropics also dominate global biodiversity, and have frequently been considered stable, safe havens for fauna and flora compared with the more variable high latitudes. Finally, because the global atmosphere–ocean circulation is driven by temperature gradients, tropical temperatures provide a linchpin on which the rest of the general circulation depends.

In the past few years, studies[12–14] based on new temperature proxy measurements, and on better-preserved records from established proxies, have produced warmer estimates for the tropics (5–10 °C warmer than modern values) than those drawn from previous work. Debate continues about whether earlier estimates were systematically biased to cool values. One independent line of evidence for an unchanging tropical climate comes from terrestrial palaeotemperature proxies derived from leaf shape, which support the idea that tropical temperatures were near modern values (24–26 °C)[7]. Of course, leaf-derived tropical temperatures could be wrong as well. Head and colleagues[6] show that this may be the case.

In the Cerrejón Formation of Colombia, South America, Head *et al.* have discovered fossil vertebrae aged between 58 million and 60 million years old, estimated to be from eight individuals of the largest species of snake ever found. But what makes the study so intriguing is that the authors relate the animal's immense projected size — 13 metres long and more than 1 tonne in weight — to a minimum annual mean temperature. To do this, they use an empirical relationship between temperature and size derived from modern organisms. The method has a biophysical grounding in the metabolism of large, air-breathing, terrestrial poikilotherms (animals whose internal temperature varies with ambient temperature). Essentially, poikilothermic animals must sustain a minimal metabolism to survive and, making the standard assumption that this metabolic rate scales with temperature, larger poikilotherms must live in warmer environments. This is the case with snakes today. Head *et al.* estimate that the giant snake required minimum temperatures of 32–33 °C, which is 6–8 °C warmer than temperatures reconstructed from floras within the same formation, and much warmer than modern values.

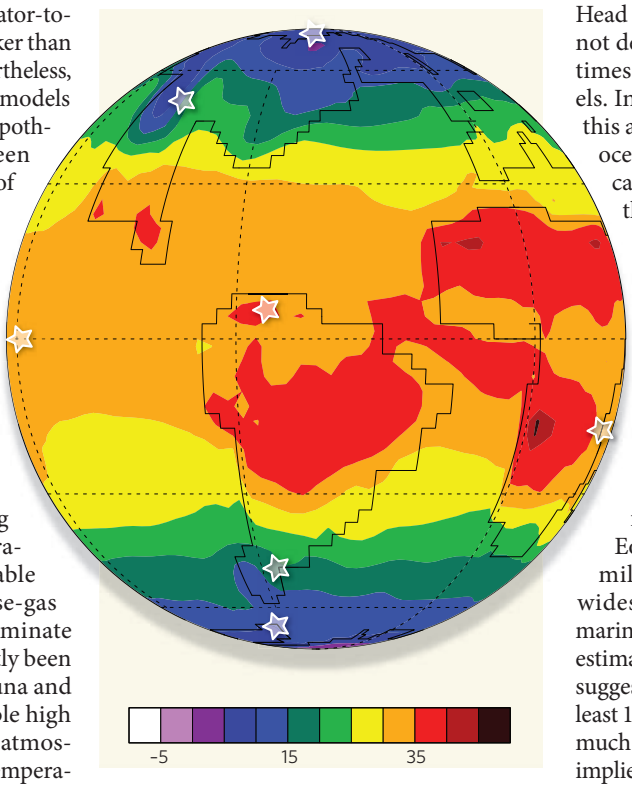


Figure 1 | Simulation of annual average surface temperatures about 58 million years ago. Stars indicate the localities for which temperature estimates exist with ages close to those of Head and colleagues' discoveries[6] in the Cerrejón Formation. In each locality, simulated temperatures match well with those estimated from temperature proxy estimates, suggesting that, with reasonable atmospheric concentrations of greenhouse gases, current models can simulate climate at this time. The simulation was carried out with the National Center for Atmospheric Research Community Climate System model (version 3), with boundary conditions for the early Palaeogene and an atmospheric CO_2 concentration of 2,240 parts per million. Temperature proxy reconstructions are from refs 3–6,12 and 14. Sea surface temperatures derived from oxygen isotopes in planktonic foraminifera that have not been proven to be as well preserved as those reported in refs 12 and 13 have been omitted from the comparison.

If we assume that Head and colleagues' temperature estimates are accurate, and it is a big assumption, there are major implications. First, there is no tropical thermostat: although negative feedbacks may slow or inhibit tropical warming, they do not provide a hard limit, and theories that predict the existence of thermostats[11] are invalid. Second, by comparing their snake-derived estimate with a palaeotemperature estimate from high-latitude Patagonia, Head *et al.* find that temperature gradients did not depart markedly from those of modern times, verifying the results of climate models. Indeed, temperatures reconstructed for this age can now be reproduced by coupled ocean–atmosphere models, provided tropical temperatures are as hot as indicated by the new results[6] (Fig. 1).

Third, although the flora and fauna in the Cerrejón Formation were remarkably resilient, and thrived in apparently hotter and wetter conditions than those of any modern rainforest setting, they may have lived near the limit of their tolerance. As Head *et al.*[6] remark, the span of time between 58 million and 60 million years ago was cooler than subsequent intervals, and further warming during the Palaeocene–Eocene Thermal Maximum, around 55.5 million years ago, could have produced widespread heat-death in terrestrial and marine ecosystems[14]. Finally, new temperature estimates from a multiple proxy approach[6,12–14] suggest that global mean temperatures were at least 10 °C warmer than modern temperatures, much warmer than previously estimated. That implies either that global average temperatures were very sensitive to greenhouse-gas forcing, or that concentrations of greenhouse gases were at the upper end of their reconstructed range[15].

All that said, these implications are based on a new type of proxy: Head and colleagues' findings are the result of probably the first study in 'snake palaeothermometry', and as such must be viewed with caution. Is the empirical link between size and temperature really generalizable and accurate? Could the ability to lose heat be an important limitation for these giant snakes, rendering Head and colleagues' extrapolations moot? Can a few vertebrae truly provide accurate estimates of snake size? Why have similarly giant snakes not been found in other warm intervals?

The findings attest to the resiliency of tropical ecosystems in the face of extreme warming, but more work is clearly necessary. For the moment, however, the burden of proof is on those who argue that the tropics do not warm substantially in a greenhouse world. ■

Matthew Huber is in the Department of Earth and Atmospheric Sciences and the Purdue Climate Change Research Center, 550 Stadium Mall Drive, Purdue University, West Lafayette, Indiana 47907, USA.
e-mail: huberm@purdue.edu

1. Markwick, P. J. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **137,** 205–271 (1998).
2. Greenwood, D. R. & Wing, S. L. *Geology* **23,** 1044–1048 (1995).
3. Tripati, A., Zachos, J., Marincovich, L. Jr & Bice, K. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **170,** 101–113 (2001).
4. Wilf, P. *GSA Bull.* **112,** 292–307 (2000).
5. Poole, I., Cantrill, D. & Utescher, T. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **222,** 95–121 (2005).
6. Head, J. J. *et al. Nature* **457,** 715–717 (2009).
7. Shackleton, N. & Boersma, A. *J. Geol. Soc.* **138,** 153–157 (1981).

8. Crowley, T. & Zachos, J. in *Warm Climates in Earth History* (eds Huber, B., MacLeod, K. & Wing, S.) 50–76 (Cambridge Univ. Press, 2000).
9. Huber, M. & Sloan, L. C. *Geophys. Res. Lett.* **28,** 3481–3484 (2001).
10. Pierrehumbert, R. T. *J. Atmos. Sci.* **52,** 1784–1806 (1995).
11. Lindzen, R. S., Chou, M. D. & Hou, A. Y. *Bull. Am. Meteorol. Soc.* **82,** 417–432 (2001).
12. Pearson, P. N. *et al. Geology* doi:10.1130/G23175A.1 (2007).
13. Norris, R. D., Bice, K. L., Magno, E. A. & Wilson, P. A. *Geology* **30,** 299–302 (2002).
14. Huber, M. *Science* **321,** 353–354 (2008).
15. Zachos, J. C., Dickens, G. R. & Zeebe, R. E. *Nature* **451,** 279–293 (2008).

QUANTUM OPTICS

A shift on a chip

Douglas H. Bradshaw and Peter W. Milonni

The Lamb shift, a minute change in certain energy levels of quantum systems that was first measured in atomic hydrogen some 60 years ago, has now been observed in a solid-state superconducting system.

The emission and absorption of light by atoms can be significantly affected by their environment. For many years, physicists have studied how atoms behave in cavities that confine light and restrict the frequencies with which the atoms interact. Cavity quantum electrodynamics (cavity QED) experiments, which examine how light and matter interact in a cavity, can be designed such that an atom is well described as a two-state system (or 'qubit') interacting with a single light frequency in a nearly lossless cavity. It has been observed, for instance, that the frequency of the light emitted (or absorbed) in an atomic transition can be altered by a cavity. More recently, similar effects have been observed in circuit QED, in which pieces of solid-state superconducting systems acting as qubits are embedded in on-chip circuits that are, in effect, one-dimensional cavities[1,2]. Writing in *Science*, Fragner *et al.*[3] now report experiments in which one of the most studied effects of QED in atomic physics — the Lamb shift — has been measured in circuit QED (Fig. 1).

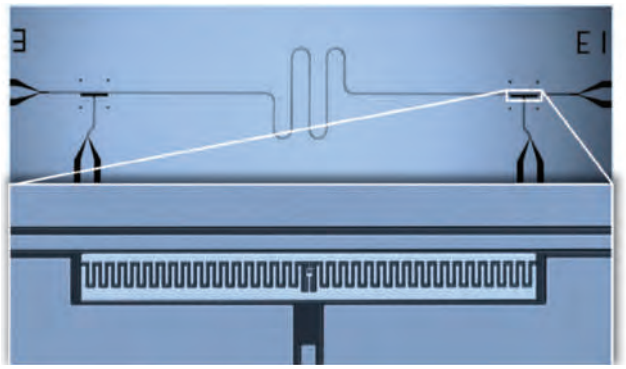


Figure 1 | Lamb shift on a solid qubit. The image shows the resonator (top) used in the experiments of Fragner *et al.*[3] to detect a tiny shift in the transition frequency — the Lamb shift — of a solid-state superconducting qubit (bottom). The qubit of dimensions 0.3 mm × 30 μm is embedded in the resonator at the position indicated by the boxed area (top right). (Image taken from ref. 1.)

The Lamb shift in atomic hydrogen was famously measured some 60 years ago[4]. Experiments showed that, in a vacuum, one of the atom's energy levels is shifted very slightly from the value predicted when the effect on the electron of the electromagnetic vacuum is ignored. The corresponding shift in the frequency of the transition of the electron to the ground state, relative to the unshifted frequency (ν), is only about 4×10^{-7}. This shift can be attributed largely to the interaction of the hydrogen atom with a continuum of electromagnetic frequencies, all in the vacuum state. Quantum fluctuations of this vacuum field, associated with the emission and absorption of 'virtual' photons, cause the electron to undergo fluctuations that change its energy level from that predicted when it is assumed to interact only with the nucleus.

But cavity QED has allowed for conceptually simpler experiments in which atoms interact with only one electromagnetic-field frequency. If this frequency is exactly tuned to the atomic resonance, a quantum of energy can flow back and forth between an atom and the electromagnetic field at a rate known as the Rabi frequency (Ω). By introducing a 'detuning' (Δ) between the atomic-transition frequency and the field frequency, one can change the nature of the atom–field interaction. When the ratio Δ/Ω is large, the observable effect of the field on the atom is to shift the atomic-transition frequency rather than to cause energy to oscillate to and fro between the atom and the field. The shift is proportional to $(q+1/2)/\Delta$, where q is the number of photons in the cavity. The Lamb shift occurs when the cavity is devoid of photons ($q=0$), and is thus associated with quantum fluctuations of the vacuum state of the field. This Lamb shift for a two-state atom was first measured in a cavity QED experiment[5]; Lamb shifts amounting to about $10^{-8}\,\nu$ were measured for the smallest detunings.

In circuit QED, a qubit is a superconducting two-state system based on the Josephson junction — two superconductors separated by a thin insulator across which electron pairs can tunnel. In the very simplest approximation, two parallel junctions can form a qubit with a transition frequency controllable by a magnetic field. The cavity resonator in circuit QED is effectively a one-dimensional waveguide formed by a superconducting structure patterned on a silicon chip. An electromagnetic field in such a resonator induces transitions in a qubit inside it if its frequency is close to the qubit transition frequency. Then the system can be described in much the same way as a qubit interacting with a single field frequency in cavity QED.

In their experiments, Fragner and colleagues[3] measured ν and the qubit–field coupling constant, which describes the strength of the interaction and thus determines the Lamb shift. They then determined the Lamb shift from the difference between ν and the measured, shifted qubit transition frequency. The detuning was varied by changing the magnetic flux through the qubit circuit. For the largest detunings, the authors obtained an excellent fit of the measured Lamb shifts to the simplified theoretical predictions based on the two-state model of the parallel Josephson junctions; a more accurate theory that accounts for deviations of the Josephson pair from a two-state system gave an excellent fit for all detunings.

A notable difference between these Lamb shifts and those in cavity QED is their magnitude — approximately $0.014\,\nu$ at the smallest detunings. These relatively large shifts reflect a strong qubit–field interaction resulting from the large electric dipole moment characterizing the qubit as well as the large vacuum-field strengths possible in the micrometre-scale resonators used in circuit QED. The strong coupling (large Rabi frequency) inferred from the Lamb-shift experiments directly illustrates one reason for the growing interest in circuit QED in connection with quantum computing[6–8], which requires that information between a photon and a qubit be exchanged rapidly compared with the rates at which any other effects, such as the escape of the photon from the resonator, cause information about the qubit state to be lost. ■

Douglas H. Bradshaw and Peter W. Milonni are at the Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA.
e-mail: pwm@lanl.gov

1. Wallraff, A. *et al. Nature* **431,** 162–167 (2004).
2. Fink, J. M. *et al. Nature* **454,** 315–318 (2008).
3. Fragner, A. *et al. Science* **322,** 1357–1360 (2008).
4. Lamb, W. E. & Retherford, R. C. *Phys. Rev.* **72,** 241–243 (1947).
5. Brune, M. *et al. Phys. Rev. Lett.* **72,** 3339–3342 (1994).
6. Makhlin, Y., Schön, G. & Shnirman, A. *Rev. Mod. Phys.* **73,** 357–400 (2001).
7. Blais, A., Huang, R.-S., Wallraff, A., Girvin, S. M. & Schoelkopf, R. J. *Phys. Rev.* A **69,** 062320 (2004).
8. Schoelkopf, R. J. & Girvin, S. M. *Nature* **451,** 664–669 (2008).

NATURE|Vol 457|5 February 2009

BIOGEOCHEMISTRY

Early animals out in the cold

Jochen J. Brocks and Nicholas J. Butterfield

The enduring controversy about the appearance of animals in the evolutionary record takes a fresh twist with an analysis of molecular fossils that places the rise of the sponge lineage before 635 million years ago.

Charles Darwin was famously sceptical about the sudden appearance of fully formed animals (metazoans) in the Early Cambrian fossil record, beginning some 542 million years ago. To a degree, he has been vindicated by the discovery of animal and animal-like fossils extending throughout the preceding Ediacaran Period, which followed the end of the second of the great Cryogenian ice ages around 635 million years ago (Fig. 1). But there the trail runs out. So is this really where metazoan life began? Or is it merely the point at which a capricious fossil record disappears?

In the absence of shells or bones, the fossil record of animals can fade away to localized snapshots, such as the remarkable diversity of early animal-like fossils in the Doushantuo biota of southern China[1]. However, estimates of evolutionary first appearance require a fundamentally more reliable type of data[2]. This is where Gordon Love and colleagues[3] (page 718 of this issue) check in with their analysis of fossil biomarkers — geologically robust and taxonomically distinctive hydrocarbon molecules, derived primarily from the lipid membranes of once-living organisms. And when it comes to tracking primitive animals, the key biomarker is a 30-carbon steroid called 24-isopropylcholestane (24-ipc). The only known sources of this compound are species of the Demospongiae, one of the three main classes of extant sponges (phylum Porifera).

Love *et al.*[3] focused on an unusually complete sequence of sedimentary rock in Oman. They not only document an abundance of 24-ipc throughout the Ediacaran, but also trace it into underlying Cryogenian strata — compelling evidence that the organisms producing this signal were present before the end of the 635-million-year-old glacial event (Fig. 1). Crucially, the authors show that the biomarkers could not have migrated from younger rocks. They achieved this by catalytically cracking the immobile organic matrix in the sediments, releasing 24-ipc biomarkers in similar high abundance to those in the associated soluble extracts.

This rigorous screening procedure was not used in a previous report[4] of 24-ipc from much older rocks, extending back to about 1,600 million years ago. Love *et al.*[3] attribute the conspicuously lower concentrations of the compound in that report to an alternative, non-sponge source, although this interpretation diminishes the value of 24-ipc as a taxonomically diagnostic biomarker. However, there is a strong possibility that the low concentrations of 30-carbon steroids in pre-Cryogenian rocks represent secondary contamination[5]. If so, the new data[3] provide an even crisper signal for a Cryogenian first appearance of 24-ipc and its biological source.

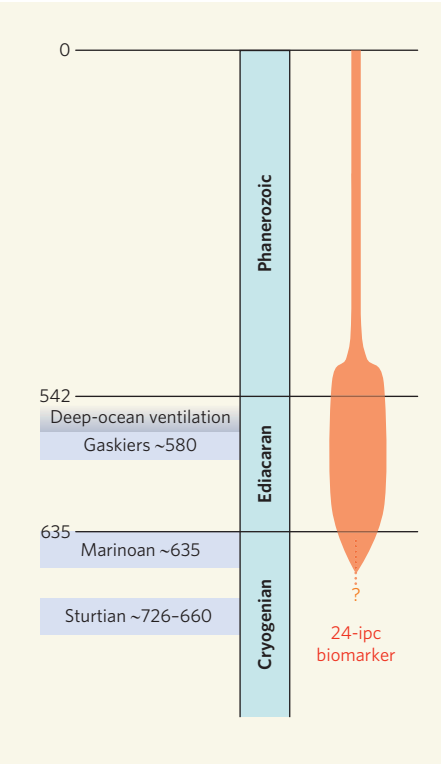


Figure 1 | Early animal evolution and the 24-isopropylcholestane (24-ipc) biomarker. The Cryogenian and Ediacaran were an interval of great environmental upheaval, including severe glaciations (the Sturtian and Marinoan, which may have been global in extent, and the Gaskiers); extreme perturbations of the global carbon cycle; and ventilation of the deep oceans with oxygen. All of these events have been cited as potential triggers for the origin of animals. Love *et al.*[3] add to the evidence of early animal life with their identification of 24-ipc in rocks older than 635 million years. The thickness of the orange line indicates the relative abundance of the 24-ipc biomarker at different times. Numbers indicate time in millions of years ago; not to scale.

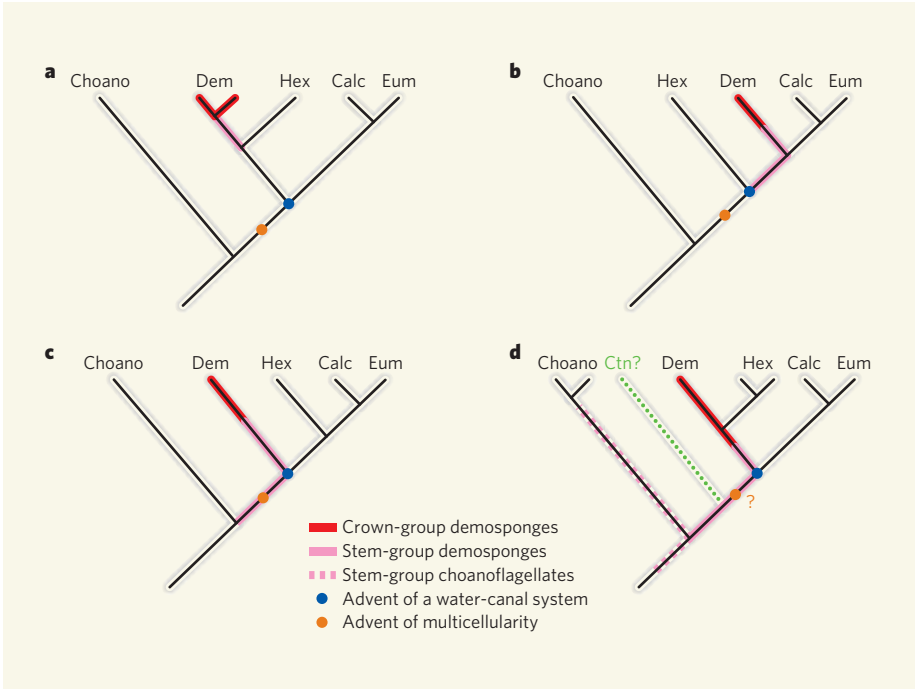


Figure 2 | Biosynthesis of the 24-ipc precursor in different evolutionary schemes. The Ediacaran and Cryogenian occurrences of 24-ipc are probably derived from stem-group demosponges (pink lines); these may or may not represent the true sponges, which exhibit multicellularity and a water-canal system. Indeed, there is no reason to rule out stem-group choanoflagellates now that these unicellular organisms are not considered directly ancestral to sponges[13]. **a**, **b**, 24-ipc is limited to true sponges. **c**, **d**, These schemes allow the possibility of (now-extinct) pre-sponge, non-metazoan organisms as the source; the phylogeny in (**d**) is currently the most strongly supported account of poriferan relationships at the class level[9]. The most recent molecular phylogeny[10] suggests that the most primitive animals are not sponges but ctenophores, a group of animals that superficially resemble cnidarian jellyfish, but that belong to a separate phylum (green dotted line in **d**). Choano, Choanoflagellida; Dem, Demospongiae; Hex, Hexactinellida; Calc, Calcarea; Eum, Eumetazoa; Ctn, Ctenophora.

So, what exactly were the organisms that produced these biomarkers? The most obvious answer, and the one that the authors[3] plump for, is that demosponges had evolved and become ecologically prominent by at least the late Cryogenian. But this conclusion overlooks the evolutionary nature of biological taxa and the incremental assembly of defining characteristics along (now-extinct) 'stem lineages' (Fig. 2). It is only with a full complement of such characteristics — in the last common ancestor of the extant 'crown group' — that modern taxonomic boundaries apply[6]. It is certainly possible, perhaps even likely, that the biomarkers from Oman reflect the existence of true, multicellular sponges with a water-canal system. But this conclusion depends on the evolutionary relationships between extant sponges (represented principally by the demosponges, hexactinellids and calcareans) and their adjacent sister groups (the single-celled choanoflagellates, and the eumetazoans; this latter group includes all metazoans apart from sponges).

A defining characteristic of crown-group demosponges, hexactinellids and calcareans is widely understood to be the development of mineralized skeletal structures, or spicules. The absence of convincing spicules in the Ediacaran or Cryogenian fossil record[7] implies that the modern poriferan classes were not fully defined until the Cambrian — and even then, seemingly bizarre combinations of spicule characteristics in Middle Cambrian fossils[8] suggest a delayed arrival of poriferan crown groups. Assuming that pre-Cambrian 24-ipc biomarkers originated from a sponge stem-group certainly does not rule out their derivation from a true sponge, and some evolutionary scenarios for the distribution of 24-ipc yield this as a unique solution (Fig. 2a,b). Other interpretations, however — including that from the most recent and comprehensively sampled analysis of hexactinellid relationships[9] — allow the biomarker biosynthesis to extend back into stem-group forms that were not sponges, and potentially not even multicellular (Fig. 2c,d). Such a possibility has important implications for the ecological interpretation of 24-ipc and the way it is applied to molecular clocks.

Despite the ambiguities, Love and colleagues' positive identification of 24-ipc in the late Cryogenian marks a considerable advance in resolving early animal evolution — particularly in light of the latest and most comprehensive molecular analysis of metazoan relationships, which no longer identifies sponges as the most primitive living animals[10] (Fig. 2d). The next steps are to find out how far back the signal can be traced in time, and how to interpret negative results. There are currently fewer than half a dozen reports of convincing biomarker occurrences of Cryogenian age, and the conspicuously low abundances of 24-ipc in post-Cambrian sediments stands at odds with the proliferation of presumed demosponge reefs in succeeding periods of Earth history.

Further sampling of the Cryogenian is clearly in order, but so too is the search for independent proxies of early animal life. Like the first predatory animals in the Ediacaran, which seem to have induced a fundamental shift in both organismal morphology and evolutionary dynamics[11], stem-group sponges may have left an indirect ecological fingerprint. It is possible, for example, that the novel feeding habits of sponges — based on the circulation of sea water through a sophisticated water-canal system — may have impinged sufficiently on the marine carbon cycle to register in the biogeochemical record[12]. Combined with new biomarker data and molecular phylogenomics, the identification of such signals promises to pinpoint the first appearance of our earliest animal ancestors. ■

Jochen J. Brocks is at the Research School of Earth Sciences, and the Centre for Macroevolution and Macroecology, the Australian National University, Canberra, 0200 ACT, Australia. Nicholas J. Butterfield is in the Department of Earth Sciences, University of Cambridge, Cambridge CB2 3EQ, UK.
e-mails: jochen.brocks@anu.edu.au; njb1005@cam.ac.uk

1. Xiao, S. & Laflamme, M. *Trends Ecol. Evol.* **24,** 31–40 (2009).
2. Butterfield, N. J. *Integr. Comp. Biol.* **43,** 166–177 (2003).
3. Love, G. D. *et al. Nature* **457,** 718–721 (2009).
4. McCaffrey, M. A. *et al. Geochim. Cosmochim. Acta* **58,** 529–532 (1994).
5. Brocks, J. J., Grosjean, E. & Logan, G. A. *Geochim. Cosmochim. Acta* **72,** 871–888 (2008).
6. Budd, G. *Nature* **412,** 487 (2001).
7. Xiao, S., Hu, J., Yuan, X., Parsley, R. L. & Cao, R. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **220,** 89–117 (2005).
8. Botting, J. P. & Butterfield, N. J. *Proc. Natl Acad. Sci. USA* **102,** 1554–1559 (2005).
9. Dohrmann, M., Janussen, D., Reitner, J., Collins, A. G. & Wörheide, G. *Syst. Biol.* **57,** 388–405 (2008).
10. Dunn, C. W. *et al. Nature* **452,** 745–749 (2008).
11. Peterson, K. J. & Butterfield, N. J. *Proc. Natl Acad. Sci. USA* **102,** 9547–9552 (2005).
12. Sperling, E. A., Pisani, D. & Peterson, K. J. in *The Rise and Fall of the Ediacaran Biota* (eds Vickers-Rich, P. & Komarower, P.) 355–368 (Geol. Soc. Lond., 2007).
13. Carr, M., Leadbeater, B. S. C., Hassan, R., Nelson, M. & Baldauf, S. L. *Proc. Natl Acad. Sci. USA* **105,** 16641–16646 (2008).

COMPUTATIONAL CHEMISTRY

Dances with hydrogen cations

Sotiris S. Xantheas

Life depends on the flow of hydrogen cations in water, yet their dynamic behaviour when in complex with water molecules is unknown. The latest computer simulations cast light on the jiggling of these hydrated ions.

In water, hydrogen cations (H^+) abound, but they exist only as complexes with water molecules. One of the most important of these complexes is the Zundel cation, in which a hydrogen cation is shared by two water molecules. The structure of the Zundel cation has been known for years, owing to evidence from infrared (IR) spectra. But its dynamic behaviour — how the hydrogen cation moves between the two water molecules — is unknown. In *Angewandte Chemie*, Vendrell *et al.*[1] report accurate computer simulations of the IR spectrum of the Zundel ion in the gas phase, and of analogues in which hydrogen atoms have been replaced with deuterium atoms. This allows the first complete characterization of the complex molecular vibrations of Zundel ions, providing information that might contribute to a long-sought-after goal — an accurate computational model of how hydrogen ions are transported through liquid water.

Hydrogen cations are ubiquitous in nature, and are vital components of many chemical and biological environments. For example, they take part in acid–base reactions that determine the formation, fate and transport of the main environmental pollutants that cause acid rain; they are pumped across cell membranes by dedicated proteins, creating gradients in pH and charge that act as energy reservoirs for the cell; and hydrogen-ion movement, when coupled to electron transfer in enzymes, allows bioenergetic conversions to occur, and 'activates' enzyme substrates, readying them to take part in catalytic bond-breaking and bond-making reactions.

Curiously, hydrogen cations seem to diffuse faster through water than do other atomic cations. In fact, hydrogen-cation 'diffusion' in water involves the concerted making and breaking of many bonds in networks of water molecules, in a process known as the Grotthuss mechanism[2] (Fig. 1a). When a hydrogen cation forms a bond to a water molecule, other covalent and hydrogen bonds throughout the network break and re-form until a different hydrogen ion is ejected. Hydrogen cations in water are thus hydrated: they either exist in complex with individual water molecules, forming Eigen ions[3] (H_3O^+), or are shared equally by two water molecules to form Zundel ions[4] ($H_2O–H–H_2O)^+$. The exact form taken by hydrated hydrogen ions — known collectively as hydronium ions — has long received much attention[5–7].

Aqueous hydronium clusters are considered to be effective vehicles for probing the dynamic environment of hydrogen cations in more complex systems such as liquid water[8,9]. The IR spectra of hydronium ions (and of

NATURE|Vol 457|5 February 2009

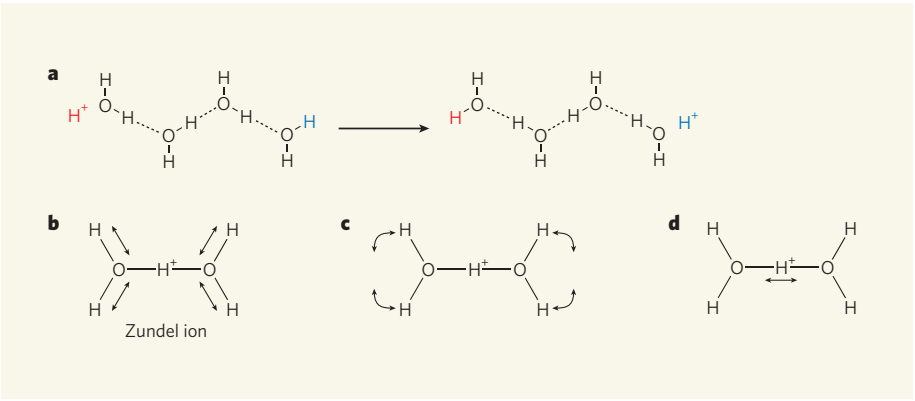


Figure 1 | The Zundel cation and its vibrations. **a**, Hydrogen ions (H^+) move through water by the Grotthuss mechanism, in which hydrogen bonds (dashed lines) and covalent bonds (solid lines) between water molecules are broken and re-formed. The mechanism can involve any number of water molecules, but only four are shown here for simplicity. **b**–**d**, The Zundel cation is the smallest structural unit that enables the sharing of a hydrogen ion by two water molecules. Its modes of vibration include stretches of the outer oxygen–hydrogen bonds (**b**), internal bending of the water molecules (**c**), and vibrations between the water molecules and the hydrogen ion (**d**). Vendrell *et al.*[1] compute the infrared spectra of Zundel cations, and of analogues of the ions in which one or more hydrogen atoms have been replaced with deuterium atoms. The different degrees of coupling between vibrational modes in each analogue produce complex spectra, which reflect the different dynamics of each analogue.

their deuterium-containing analogues) can be thought of as 'fingerprints' of the underlying molecular structure and of the dynamics of the ions' hydrogen-bonding network. The correlation[10,11] between the IR spectrum and the structure of each ion provides a method for identifying the ions' structures, and yields information about the coupling between the various vibrational modes of the ions.

But the dynamic behaviour and complex vibrational motions of hydronium ions are by no means obvious. The simple harmonic picture of vibrations (which assumes that atoms behave as frictionless masses connected by springs) is valid only around the equilibrium positions of the atoms; in Zundel cations, IR spectra instead suggest the existence of large-amplitude, anharmonic motions. The presence of Fermi resonances — overlapping absorption lines that arise from strong couplings between vibrational states of similar energies — in the IR spectra of Zundel cations also makes it difficult to work out the dynamic behaviour of the ions from the spectra. Several approaches[12,13] have therefore been introduced to account for the Fermi resonances and to describe large-amplitude molecular motions.

A good way to validate the theoretical approaches and to refine structural models of hydronium ions is to compute the IR spectra that would be obtained from an assumed structure, and then to compare these spectra with experimental data. Vendrell *et al.*[1] computed the IR spectra for the Zundel cation by integrating the ion's vibrational, time-dependent Schrödinger equation in 15 dimensions — one dimension for each vibrational degree of freedom — using a 'wavefunction propagation' approach[14]. This required an analytical description of the potential energy surface of the cation to describe the total energy of the system as a function of each degree of freedom. The authors thus obtained spectra for the Zundel ion itself, and for the deuterium-containing ions $D(D_2O)_2^+$, $H(D_2O)_2^+$ and $D(H_2O)_2^+$ (where D is deuterium). They found that increasing the number of deuterium ions in the Zundel ion progressively complicates the spectrum — creating a bigger "mess", to use the authors' own word.

Perhaps unsurprisingly, Vendrell and colleagues' computed spectra show that the vibrations of 'free' oxygen–hydrogen bonds — those containing hydrogen atoms that don't participate in hydrogen bonding (Fig. 1b) — are broadly unaffected when their hydrogen atoms are replaced with deuterium atoms. The only effect of deuteration is a shift of the relevant peaks towards lower frequencies, as would be expected from the change in mass associated with the substitution. By contrast, the peaks of the spectra associated with internal bending of the water molecules (Fig. 1c), and with vibrations of the bonds and atoms that form hydrogen bonds (Fig. 1d), vary dramatically with deuteration of the ions.

Vendrell and colleagues' approach allows the various peaks in the spectra to be precisely explained in terms of combinations of the constituent vibrational modes of Zundel ions. They thus show that the complicating effects of deuterium atoms on the dynamics of Zundel ions depend on which hydrogens within the cluster they replace. To be precise, the presence of deuterium in the water molecules induces strong couplings between the movements of the central hydrogen ion and the water molecules. Conversely, when the hydrogen ion is replaced with a deuterium ion, these motions are decoupled and the resulting spectra are simpler than those of non-deuterated Zundel ions.

This work[1] is the first step in obtaining a quantitative picture of the complex dynamics associated with the interactions of hydrogen cations with water. It is a great start, but there is still a long way to go. The spectra of hydrogen ions that have more water molecules[9] must now be analysed for us to understand how such spectra evolve with cluster size, and to unravel the role of the collective effects that facilitate the Grotthuss mechanism in liquid water. The potential energy surfaces of such clusters have more degrees of vibrational freedom than Zundel ions, making it more difficult both to obtain an analytical description of each cluster, and to solve the time-dependent Schrödinger equation in all dimensions. New theoretical approaches will therefore be needed to address these issues, and to deal with possible strong couplings between different vibrational modes in many dimensions, which would produce even more complex spectra. Alternatively, simpler descriptions of the underlying intermolecular interactions between hydrogen cations and water can be sought[6,15], in which case Vendrell and colleagues' approach will provide the machinery to fit those descriptions so that they can reproduce experimental spectra.

Obtaining an accurate description of the interactions between water molecules (or between water molecules and ions), and understanding the collective physical phenomena that are present in water, will result in better models of the liquid that can be used to study solvation and reactions in aqueous environments. This will ultimately offer molecular-level insight into important environmental processes — such as the fate and transport of contaminants in rivers and aquifers — and decipher the function of water in confined spaces of biological interest. ■

Sotiris S. Xantheas is in the Chemical and Materials Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, USA.
e-mail: sotiris.xantheas@pnl.gov

1. Vendrell, O., Gatti, F. & Meyer, H.-D. *Angew. Chem. Int. Edn* **48,** 352–355 (2009).
2. de Grotthusss, C. J. T. *Ann. Chim.* **58,** 54–73 (1806).
3. Eigen, M. *Angew. Chem. Int. Edn* **3,** 1–19 (1964).
4. Zundel, G. & Metzer, H. *Z. Phys. Chem.* **58,** 225–245 (1968).
5. Agmon, N. *Chem. Phys. Lett.* **244,** 456–462 (1995).
6. Schmitt, U. W. & Voth, G. A. *J. Chem. Phys.* **111,** 9361–9381 (1999).
7. Smondyrev, A. M. & Voth, G. A. *Biophys. J.* **82,** 1460–1468 (2002).
8. Asmis, K. R. *et al. Science* **299,** 1375–1377 (2003).
9. Headrick, J. M. *et al. Science* **308,** 1765–1769 (2005).
10. Cabarcos, O. M., Weinheimer, C. J., Lisy, J. M. & Xantheas, S. S. *J. Chem. Phys.* **110,** 5–8 (1999).
11. Nauta, K. & Miller, R. E. *Science* **287,** 293–295 (2000).
12. Bowman, J. M. *J. Chem. Phys.* **68,** 608–610 (1978).
13. Gerber, R. B. & Ratner, M. A. *Chem. Phys. Lett.* **68,** 195–198 (1979).
14. Beck, M. H., Jäckle, A., Worth, G. A. & Meyer, H.-D. *Phys. Rep.* **324,** 1–105 (2000).
15. Hodges, M. P. & Stone, A. J. *J. Chem. Phys.* **110,** 6766–6772 (1999).

Q&A

NEUROSCIENCE

Glia — more than just brain glue

Nicola J. Allen and Ben A. Barres

Glia make up most of the cells in the brain, yet until recently they were believed to have only a passive, supporting role. It is now becoming increasingly clear that these cells have other functions: they make crucial contributions to the formation, operation and adaptation of neural circuitry.

How do glia differ from neurons?

The defining characteristic of a neuron is its ability to transmit rapid electrical signals in the form of action potentials. All other neural cells that lack this property are categorized into a broad class termed glia. Neurons are arranged in networks (circuits), and communicate with each other via specialized intercellular adhesion sites called synapses. Neuronal signalling involves the propagation of an action potential down a neuron's axonal process to a presynaptic terminal; the depolarization of the terminal and release of neurotransmitters; binding of the released neurotransmitters to receptors on the postsynaptic membrane of another neuron; and the subsequent depolarization of this second neuron, propagating the signal further. Glia do not fire action potentials, but instead surround and ensheath neuronal cell bodies, axons and synapses throughout the nervous system.

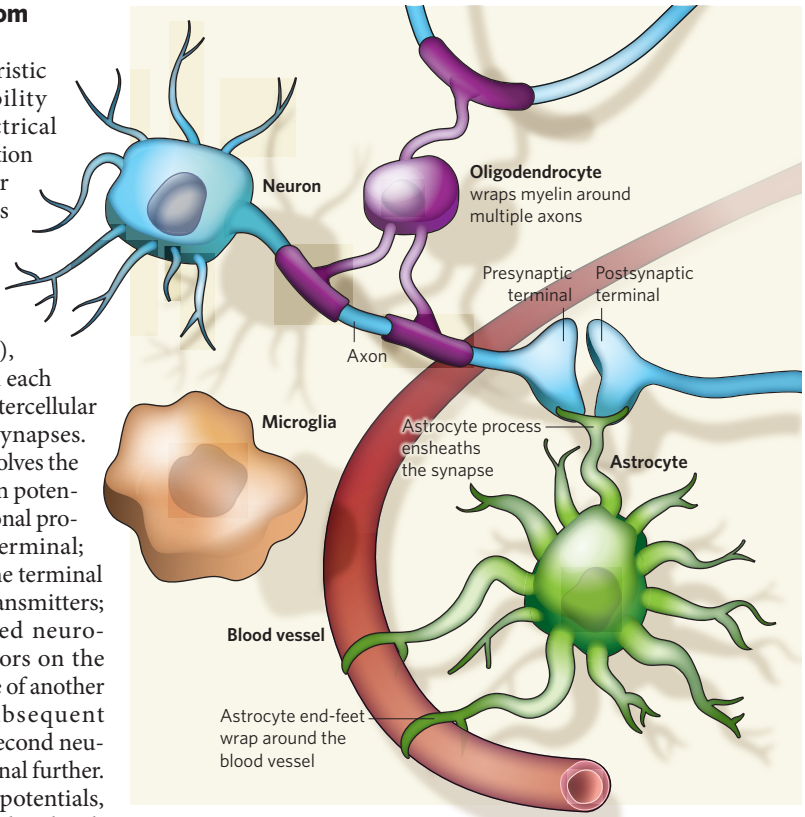


Figure 1 | Glia–neuron interactions. Different types of glia interact with neurons and the surrounding blood vessels. Oligodendrocytes wrap myelin around axons to speed up neuronal transmission. Astrocytes extend processes that ensheath blood vessels and synapses. Microglia keep the brain under surveillance for damage or infection.

Are all glia the same?

No. On the basis of morphology, function and location in the nervous system, there are several classes of glia. In mammals, for example, glia are classified as microglia, astrocytes and the related Schwann cells and oligodendrocytes (Fig. 1).

Where do they originate from?

Glia and neurons mainly share a common origin — precursor cells derived from the embryonic germ layer known as the neuroectoderm. A notable exception is microglia, which are part of the immune system and enter the brain from the blood circulation early in an organism's development.

What is known about the evolution of glia?

Glia are evolutionarily conserved, being present in one form or another in most species examined, from the simplest invertebrates to humans. The proportion of glia seems to be correlated with an animal's size: the tiny nematode worm has only a few glia; some 25% of the fruitfly brain consists of glia; the mouse brain has roughly 65% of these cells; the human brain has about 90%; and the elephant brain consists of some 97% glia. As animals have evolved, glia have become not only more diverse and specialized, but also essential: without them neurons die. Furthermore, astrocytes in the human cerebral cortex are much more complex than those of other mammals, and are thought to be involved in information processing.

So what exactly do glia do?

Lots of things. The traditional view has been that glia look after neurons and maintain their proper functioning, having a somewhat passive role themselves. Established functions of glia include supporting neurotransmission, maintaining ionic balance in the extracellular space, and insulating axons to speed up electrical communication. But emerging research suggests that glia, particularly astrocytes, also have an active role in brain function and information processing — both during development and in adulthood.

What is the specific function of microglia?

These resident immune cells of the nervous system survey the brain for damage and infection, engulfing dead cells and debris. Microglia have also been implicated in synaptic remodelling during the development of the nervous system, when they are proposed to remove inappropriate synaptic connections through the process of phagocytosis. Moreover, they are activated in many neurodegenerative diseases, but whether they are helpful or harmful in these conditions is a matter of debate.

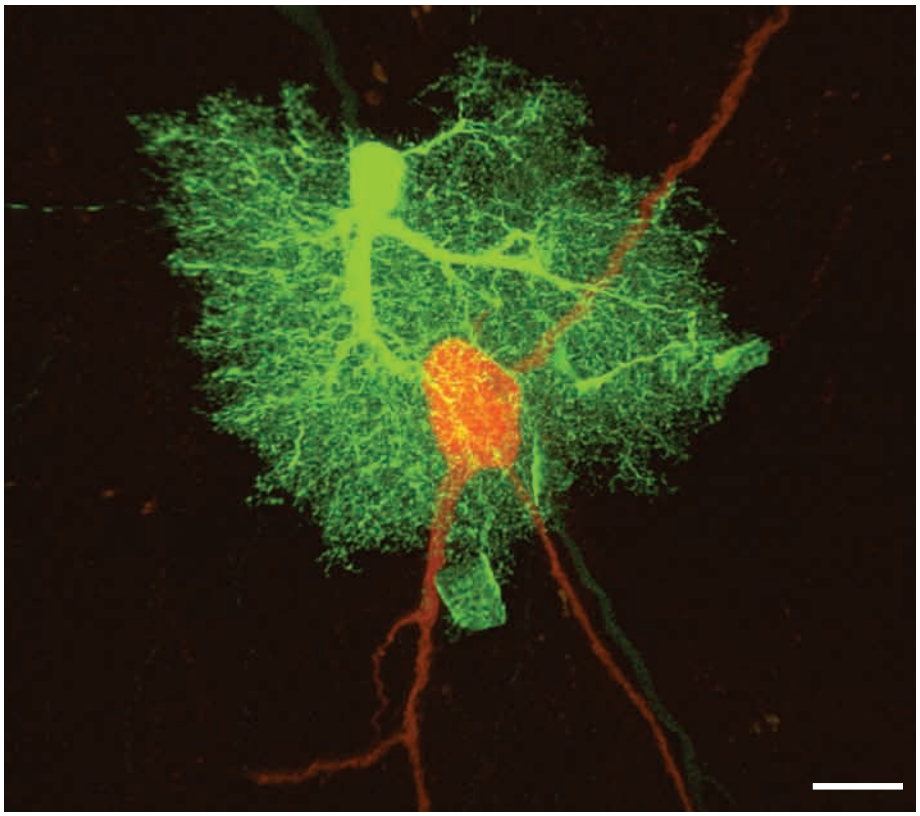


Figure 2 | An astrocyte in action. This micrograph shows a protoplasmic astrocyte (green) enveloping the cell body and the processes of a neuron (red). The bushy nature of astrocytes, evident in this image, allows them to form distinct domains in the brain. Scale bar, 10 μm. (Image courtesy of M. Ellisman and E. Bushong, Univ. California, San Diego.)

What do oligodendrocytes and Schwann cells do?

In vertebrates, these cells are essential for rapid electrical communication between neurons and their targets. Oligodendrocytes (in the central nervous system) and Schwann cells (in the peripheral nervous system) produce a lipid-rich membrane called myelin, which enwraps axons, thereby speeding up the conduction of electrical impulses. In the absence of myelin, the conduction velocity of an action potential is directly proportional to the diameter of the axon. This means that the final size of an animal would be limited by the fact that its axons would eventually become prohibitively large. The evolution of myelin has allowed animal size to increase without a corresponding increase in axon diameter, enabling rapid thought and action. Besides, myelination induces clustering of ion channels, thereby further enhancing conduction velocity. 'Demyelination' — due to damage to oligodendrocytes and Schwann cells — leads to various diseases, including multiple sclerosis and hereditary sensorimotor neuropathy.

And what about astrocytes?

Put simply, astrocytes allow neurons to function (Fig. 2). They contribute to homeostasis in the brain by providing neurons with energy and substrates for neurotransmission. They act as physical barriers between the synaptic connections of neighbouring neurons, and remove excess neurotransmitter molecules from the extracellular space, allowing discrete and precise encoding of synaptic signals and neurotransmission. Recently, unexpected roles for astrocytes have been identified — they seem to be involved in the formation of synapses and in modulating synaptic function through bidirectional communication with neurons. For this reason, the next few questions are dedicated to this, currently the 'hottest' type of glia.

How exactly do astrocytes contribute to homeostasis in the brain?

Astrocytes control blood flow through their numerous fine processes, which form close associations with both blood vessels and neurons. In response to enhanced neuronal activity, astrocytes signal to blood vessels about the need for regional increases in blood flow, which results in enhanced delivery of oxygen and glucose to the active brain region. Analysis of such changes in blood flow forms the basis of the study of brain function by functional magnetic resonance imaging (fMRI). Besides regulating blood flow, astrocytes ferry glucose and oxygen from blood to neurons. It is hypothesized that they convert glucose into lactate. Lactate is then exported to neurons, where it is converted to pyruvate to produce the cell's energy molecule ATP. Astrocytes are also responsible for terminating the action of neurotransmitters secreted by neurons and for mediating their recycling back to neurons in a process known as the glutamate–glutamine cycle.

Are all astrocytes the same?

No. These cells are broadly divided into two groups — protoplasmic astrocytes found in the brain's grey matter and fibrous astrocytes of the white matter. Protoplasmic astrocytes are intimately associated with neuronal cell bodies and synapses, whereas fibrous astrocytes are associated with neuronal axons. Furthermore, types of protoplasmic astrocyte differ between the various regions of grey matter; even within a single brain region, neighbouring astrocytes are probably different. This is not surprising, because, if they are to fulfil different functions, these cells must adapt to specific brain regions. Exact functional differences between the various types of astrocyte remain elusive.

Do astrocytes talk to each other?

Yes. They communicate with each other through waves of calcium ions, propagating information over large distances. Stimulation of one astrocyte can cause a calcium response in a subset of neighbouring astrocytes, with no response in other subsets, indicating the presence of distinct networks of astrocytes organized in a mosaic pattern. Although individual astrocytes occupy distinct domains, and cellular projections from neighbouring astrocytes do not overlap in the adult brain, these cells are linked together by structures in their cell bodies called gap junctions.

Do they communicate with neurons too?

Bidirectional communication does indeed occur between neurons and astrocytes. Individual astrocytes can make contact with and ensheath thousands of synapses formed between many different neurons. This means that synapses don't consist of just a pre- and postsynaptic neuronal element, but that many also have an astrocytic projection that envelops the synapse. This close spatial relationship has led to the term tripartite synapse, to acknowledge the astrocyte's contribution (Fig. 3). The synaptic localization of astrocytes means they are ideally placed to monitor — and respond to — synaptic activity. Moreover, astrocytes possess many of the same neurotransmitter receptors as neurons, and neurotransmitter release by neurons activates calcium-based signalling cascades in astrocytes. Astrocytes then release neuroactive substances, signalling back to neurons to form a feedback loop. The different types of molecule secreted by astrocytes can either inhibit or enhance overall levels of neuronal activity.

Do any types of glia receive direct neuronal input ?

Possibly. Cells expressing the proteoglycan NG2, which are thought to be oligodendrocyte

precursor cells, have been shown to receive direct synaptic signals from neurons, some of them even firing action-potential-like signals in response. The significance of this innervation is not known. Does it influence the decision of NG2-expressing cells to become oligodendrocytes, or even neurons? And could it result in the recruitment of NG2 cells into specific neural networks?

What is the role of glia in embryonic brain development?

Some glia give rise to neurons, and others guide neurons to their correct location in the nervous system — so they are essential for brain development. During embryonic development, a specialized type of glia called radial glia divide to form neural progenitor cells. Moreover, the long processes of radial glia span the cortex, providing tracks along which newly generated neurons migrate to reach their correct location. Once all neurons are in place, the processes of radial glia degenerate and they form cortical astrocytes. Besides guiding neurons to their correct location, glia provide a scaffold along which axons grow. They perform this pathfinding function through both attractive and repulsive interactions with receptors present on the axon.

And how do they contribute to the formation of neural networks?

They do so by aiding synapse formation and possibly synapse elimination. Astrocytes, for instance, induce synapse formation in several classes of neuron, both by direct contact with neurons and by secreting factors that regulate synapse formation as well as pre- and postsynaptic functions. But such actions are not restricted to astrocytes. Oligodendrocytes and Schwann cells also induce synapse formation between neurons. It is not clear how such glial signals act: do they provide a permissive environment for synapse formation at sites predetermined by neurons, or do they actively instruct neurons where to form synapses? As mentioned previously, microglia are also implicated in the removal of inappropriate synaptic connections and so in fine-scale 'sculpting' of neuronal networks.

Do glia play a part in disease?

Glia can be a help or a hindrance in disorders of the nervous system, and their malfunction has been implicated in many such diseases. For example, following spinal-cord injury, astrocytes form a glial scar that acts as a barrier to the regeneration of damaged axons. Moreover, in the neurodegenerative disease amyotrophic lateral sclerosis, astrocytes secrete a toxic factor that kills motor neurons — those involved in muscle function. And astrocytes can sometimes become cancerous, giving rise to brain tumours called gliomas. Also, as mentioned earlier, oligodendrocytes are the target of an autoimmune attack in multiple sclerosis that causes demyelination. Unexpectedly, profound loss of oligodendrocytes and myelin has been reported in clinical depression.

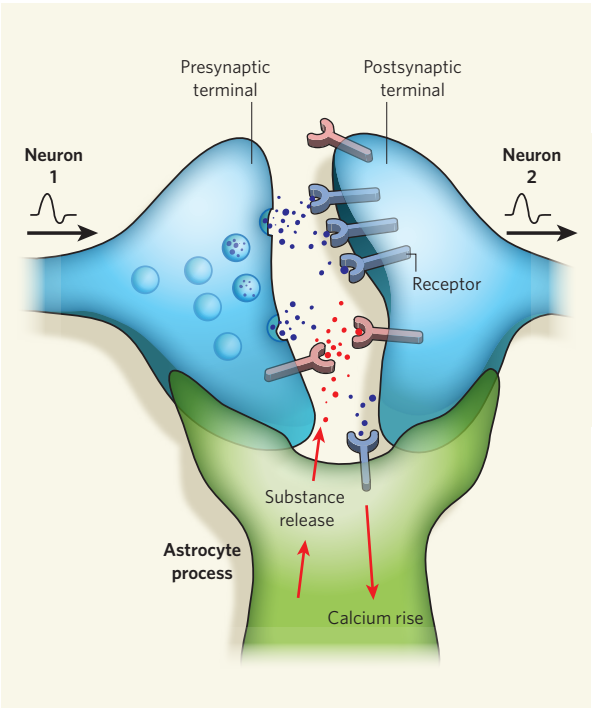


Figure 3 | A tripartite synapse. Astrocytes express many of the same receptors as neurons. When neurotransmitters are released from the presynaptic terminal of a neuron, astrocytic receptors are thought to be activated, leading to a rise in calcium ions in the astrocyte and the release of various active substances, such as ATP, which act back on neurons to either inhibit or enhance neuronal activity. Astrocytes also release proteins, which control synapse formation, regulate presynaptic function and modulate the response of the postsynaptic neuron to neurotransmitters.

What experimental models are used to study glia?

Studying the role of glia in nervous-system function is difficult because, in most organisms, glia are essential for neuronal survival and so their removal causes neuronal death. Therefore, much of what we know about glia has come from studies of isolated mammalian glia maintained *in vitro*. Although such analysis is useful and has taught us much about the basic properties of glia, it cannot tell us how glia interact with other cell types. Electrophysiological and calcium imaging studies using mammalian brain slices have begun to provide insight into both glia–neuron interaction and the role of glia in the activity of neuronal networks. Also, with advances in live imaging techniques, such as *in vivo* two-photon microscopy, glial activity and its correlation with blood flow and behaviour can be monitored in living animals. Studying small model organisms, including worms, fruitflies and fish, is another powerful approach that has allowed dissection of the role of glia in nervous-system function by means of genetic engineering.

So what is left to learn about them?

Lots! Although the recent resurgence of interest in glia has led to many exciting and unexpected discoveries about the roles of these cells in the nervous system, such findings are almost certainly just the tip of the iceberg, and there are many outstanding questions. How exactly do glia participate in the formation and functioning of neuronal networks? Do glia have essential functions beyond supporting and interacting with neurons? What is the extent of astrocytic networks? How crucial are these networks, and can they process information in the absence of neurons? How do glia contribute to disease, and might they be a potential target for drugs?

Why the resurgence of interest?

A historical difficulty with studying glia has been the 'neuro-centric' view of the brain, implicit in the name of the discipline: neuroscience. Fortunately, there is a growing appreciation of the importance of other cell types in the nervous system and their symbiotic relationship with neurons, with no single cell type now being viewed as more important than the others. By examining how all of these cells work together, neurobiologists hope to make more rapid progress in understanding how the nervous system forms, functions, adapts and can be repaired. ■

Nicola J. Allen and Ben A. Barres are in the Department of Neurobiology, Stanford University School of Medicine, Stanford, California 94305-5125, USA.
e-mails: njallen@stanford.edu; barres@stanford.edu

FURTHER READING
Allen, N. J. & Barres, B. A. Signaling between glia and neurons: focus on synaptic plasticity. *Curr. Opin. Neurobiol.* **15,** 542–548 (2005).
Barres, B. A. The mystery and magic of glia: a perspective on their roles in health and disease. *Neuron* **60,** 430–440 (2008).
Freeman, M. R. & Doherty, J. Glial cell biology in *Drosophila* and vertebrates. *Trends Neurosci.* **29,** 82–90 (2006).
Haydon, P. G. & Carmignoto, G. Astrocyte control of synaptic transmission and neurovascular coupling. *Physiol. Rev.* **86,** 1009–1031 (2006).
Kettenmann, H. & Ransom, B. R. (eds) *Neuroglia* 2nd edn (Oxford Univ. Press, 2005).
Nave, K.-A. & Trapp, B. D. Axon-glial signaling and the glial support of axon function. *Annu. Rev. Neurosci.* **31,** 535–561 (2008).
Wang, D. D. & Bordey, A. The astrocyte odyssey. *Prog. Neurobiol.* **86,** 342–367 (2008).

ARTICLES

A high-mobility electron-transporting polymer for printed transistors

He Yan[1], Zhihua Chen[1], Yan Zheng[1], Christopher Newman[1], Jordan R. Quinn[1], Florian Dötz[2], Marcel Kastler[3] & Antonio Facchetti[1]

Printed electronics is a revolutionary technology aimed at unconventional electronic device manufacture on plastic foils, and will probably rely on polymeric semiconductors for organic thin-film transistor (OTFT) fabrication. In addition to having excellent charge-transport characteristics in ambient conditions, such materials must meet other key requirements, such as chemical stability, large solubility in common solvents, and inexpensive solution and/or low-temperature processing. Furthermore, compatibility of both p-channel (hole-transporting) and n-channel (electron-transporting) semiconductors with a single combination of gate dielectric and contact materials is highly desirable to enable powerful complementary circuit technologies, where p- and n-channel OTFTs operate in concert. Polymeric complementary circuits operating in ambient conditions are currently difficult to realize: although excellent p-channel polymers are widely available, the achievement of high-performance n-channel polymers is more challenging. Here we report a highly soluble (~60 g l^{-1}) and printable n-channel polymer exhibiting unprecedented OTFT characteristics (electron mobilities up to ~0.45–0.85 $cm^2 V^{-1} s^{-1}$) under ambient conditions in combination with Au contacts and various polymeric dielectrics. Several top-gate OTFTs on plastic substrates were fabricated with the semiconductor-dielectric layers deposited by spin-coating as well as by gravure, flexographic and inkjet printing, demonstrating great processing versatility. Finally, all-printed polymeric complementary inverters (with gain 25–65) have been demonstrated.

Electronic devices performing simple operations or functions based on OTFTs (see Fig. 1 for structure and operation) offer unique attractions compared to well-established silicon electronics[1,2]. These attractions include high-throughput and inexpensive production, mechanical flexibility, light weight and efficient integration within the supply chain, as well as great opportunities for new fundamental studies[3–12]. Although device speeds may be modest when compared to inorganic-based circuits, the above-mentioned characteristics

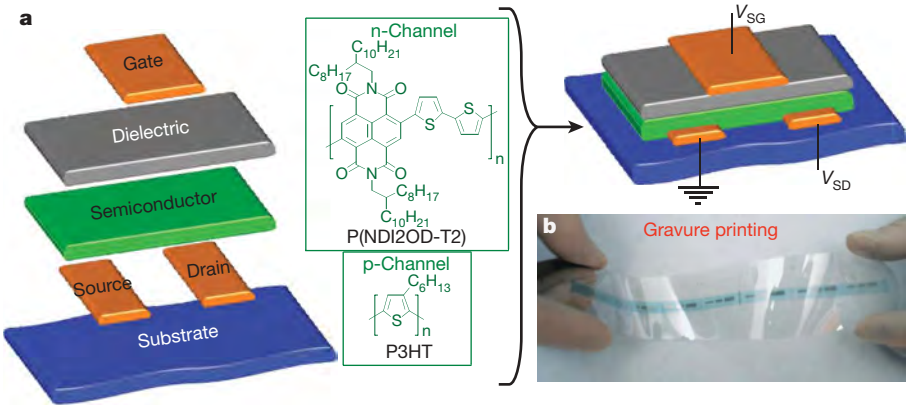


Figure 1 | Organic thin-film transistor structure, fabrication and operational principles. a, Illustration of the TFT material components (left), chemical structure of P(NDI2OD-T2) and P3HT semiconducting polymers, and of the top-gate bottom-contact (TGBC) TFT architecture (right) used in this study. For device fabrication (see Methods for details), first source and drain electrodes (Au) are fabricated by vacuum thermal evaporation on the plastic substrate (PET, polyethylene terephthalate). Next the polymeric semiconductor layer is deposited on the substrate contacts by either spin-coating or printing P(NDI2OD-T2) solutions. The gate dielectric layer is either spin-coated or gravure-printed on top of the semiconducting polymer. The device structure is completed by vapour deposition of the gate lines. In a TFT architecture, the current flow between source and drain electrodes (I_{SD}) on application of a drain–source bias (V_{SD}) is modulated by the bias applied between gate and source electrodes (V_{SG}). When $V_{SG} = 0$ V, I_{SD} is minimal and the device is in the 'off' state. When $V_{SG} \neq 0$ V is applied ($V_{SG} > 0$ V for n-channel transistors) the device turns 'on', and charge carriers are accumulated at the semiconductor–dielectric interface, resulting in a gate-controlled I_{SD}. Principal TFT figures-of-merit include the field-effect mobility (μ) and the current on-to-off ratio (I_{on}/I_{off}), defining average charge carrier drift velocity and source–drain current ratio between 'on' and 'off' states, respectively. **b**, Optical images of gravure-printed TGBC TFTs on PET before top-gate contact deposition.

[1]Polyera Corporation, 8045 Lamon Avenue, Skokie, Illinois 60077, USA. [2]BASF Global Research Center Singapore, Science Park Road 61, Singapore 112575. [3]BASF SE, GKS/E-B001, 67056 Ludwigshafen, Germany.

make this technology attractive for unconventional circuit, sensor and display applications[13–17]. To achieve these goals, semiconductor and dielectric solution-processability is the key prerequisite for inexpensive device assembly by spin-coating, casting or printing[18]. Polymeric materials are obvious candidates for OTFTs because they enable ink formulations with tuned rheological properties[18–20]. Several solution-processable and printable polymeric gate dielectrics exhibiting excellent performance have been developed[21,22]. As far as solution-processable polymeric semiconductors are concerned, impressive progress has been made in developing p-channel materials, with P3HT, F8T2, PBTTT and PQT being among the most investigated polymers (see structures in Supplementary Fig. 1)[23–25]. Thin-film transistors (TFTs) based on these polymers typically exhibit hole mobilities of $\mu_h \approx 0.01$–$1\ cm^2\ V^{-1}\ s^{-1}$ in ambient conditions. However, despite the impressive performance of a few molecular n-channel semiconductors[26–28], recent progress in developing new electron-depleted π-conjugated cores[29,30], and fundamental studies demonstrating general electron transport in polymers under selected device fabrication and measurement conditions[31], the realization of high-performance, ambient-stable n-channel polymeric semiconductors remains challenging. To the best of our knowledge, the best performing structures are the recently reported dithiophenealkylimide-based[32] and perylene-based[33,34] polymers exhibiting promising electron mobilities ($\mu_e = 0.001$–$0.01\ cm^2\ V^{-1}\ s^{-1}$, current on-to-off ratio (I_{on}/I_{off}) $\approx 10^5$ in vacuum) and the ladder-type BBL material exhibiting a μ_e of $\sim$0.001–0.1 $cm^2\ V^{-1}\ s^{-1}$ ($I_{on}/I_{off} \approx 10^1$–$10^4$ in air, film processing involves methanesulphonic acid)[35]; see structures in Supplementary Fig. 1. The discovery of high-performance, stable and readily processable electron-transporting polymeric semiconductors would represent a major step towards polymeric complementary circuit technologies[36], where the combination of p- and n-channel transistors results in far greater circuit speeds, lower power dissipation and more stable operation[18]. Polymeric complementary circuits are currently unrealized because of the absence of a suitable n-channel component. Furthermore, besides the technological aspect, fundamental questions arise as to whether n-channel polymers can ever approach p-channel polymer charge transport efficiencies in a TFT architecture, whether these devices are operable and stable in ambient conditions, and finally whether the same high-performance semiconductors exhibit sufficient solubility so that various deposition techniques can be used for film processing. In this Article we demonstrate the realization of high-mobility top-gate TFTs based on a naphthalene-bis(dicarboximide) (NDI) polymer in combination with several polymeric dielectric materials. Furthermore, thanks to the efficient electron injection into this semiconductor from high-work-function metal contacts (Au), all-polymer complementary logic is demonstrated.

Semiconductor polymer characterization

The NDI-based polymer poly{[*N*,*N*′-bis(2-octyldodecyl)-naphthalene-1,4,5,8-bis(dicarboximide)-2,6-diyl]-alt-5,5′-(2,2′-bithiophene)}, (P(NDI2OD-T2), Polyera ActivInk N2200, Fig. 1), was synthesized by reacting *N*,*N*′-dialkyl-2,6-dibromonaphthalene-1,4,5,8-bis(dicarboximide) with 5,5′-bis(trimethylstannyl)-2,2′-dithiophene (see Methods). Compared to previously utilized perylene-bis(dicarboximide) (PDI) monomers, the NDI core was identified as the key polymer building block to ensure a strong electron-depleted electronic structure and, equally important, a regioregular and highly π-conjugated polymeric backbone[33,37]. Unless indicated, all the P(NDI2OD-T2) batches used in this study were purified by Soxhlet extraction with acetone followed by multiple dissolution-precipitation ($CHCl_3$-MeOH) procedures, affording an weight-average molecular weight (M_w) of $\sim$280 kDa and a polydispersity (PD) of $\sim$5.5 (gel permeation chromatography measurements, Supplementary Fig. 2). The sample purity was assessed by elemental analysis and ^{1}H NMR spectroscopy (Supplementary Fig. 3). The room temperature solubilities of this polymer in conventional organic solvents such as xylene and dichlorobenzene (DCB) are as high as $60\ g\ l^{-1}$. Cyclic voltammetry experiments reveal two reversible reductions, with the first and second half-wave potentials ($E_{1/2}$ versus the standard calomel electrode, Supplementary Fig. 4) located at -0.49 and -0.96 V, respectively, corroborating the electron-poor nature of this π-conjugated polymer. Despite the far larger bandgap of the naphthalene- ($\sim$3 eV) versus the perylene-bis(dicarboximide) ($\sim$2 eV)[37] cores, the optical absorption spectra of P(NDI2OD-T2) (Supplementary Fig. 5) reveal an optical gap of only $\sim$1.45 eV(ref. 37). This result, enabled by the regioregular polymeric backbone, confirms the extended π-conjugated electronic structure of this semiconductor. The corresponding LUMO (lowest-unoccupied molecular orbital) energy of about -4.0 eV is among the lowest reported to date for a semiconducting polymer, approaching those of strongly electron-depleted core-cyanated rylenes[38]. Differential scanning calorimetry of P(NDI2OD-T2) exhibits no thermal transitions up to $\sim$300 °C (Supplementary Fig. 6), suggesting the absence of mesophase formation before melting. The reversibility of the endothermic and exothermic processes ($\Delta H = 108.52$ (endo), 113.74 (exo) $J\ g^{-1}$) is evidence of excellent thermal stability. Atomic force microscopy (AFM, Supplementary Fig. 7) images of P(NDI2OD-T2) thin-films spun onto Si-SiO_2 substrates and annealed over a wide temperature range (110 → 210 °C) exhibit similar fibre-like morphologies having $\sim$200–300-nm-wide lateral dimensions. Furthermore, the corresponding wide-angle X-ray diffraction (XRD) scans reveal negligible Bragg reflection intensities under all film thermal annealing conditions (Supplementary Fig. 8). The mostly amorphous nature of these P(NDI2OD-T2) films is surprising considering the large electron mobilities achievable (see below).

TFT fabrication and general performance

The semiconducting properties of P(NDI2OD-T2) were optimized in a top-gate bottom-contact (TGBC, Fig. 1) architecture having glass or PET (substrate)/Au (source-drain contacts)/P(NDI2OD-T2)/polymeric dielectric/Au (gate contact). (Here PET indicates polyethylene terephthalate.) This structure was selected because of the superior injection characteristics of typical staggered (top-gate) architectures and considering the facile channel miniaturization for bottom-contact TFTs which could lead to high-frequency circuits[15,18]. These TGBC devices were fabricated with the P(NDI2OD-T2) film deposited by spin-coating as well as by gravure, flexographic and inkjet printing, and with the dielectric layer deposited by spin-coating. Furthermore, TGBC TFTs where both the semiconductor and the dielectric layers were gravure-printed are demonstrated. All device deposition processes were performed in ambient conditions with the exception of the Au contact vapour deposition and the film drying steps ($\leq$110 °C). TFT fabrication details, measurements, and performance parameter calculations are reported in Methods. Table 1 collects the transistor performance parameters measured in ambient conditions, including the field-effect electron mobility (μ_e), current on-to-off ratio (I_{on}/I_{off}), turn-on voltage (V_{on}), threshold voltage (V_T) and subthreshold swing (S).

The polymeric dielectrics used in this study were selected to cover a variety of chemical structures, surface energies and dielectric constant (k) values, and include: CYTOP (poly(perfluoroalkenylvinyl ether), $k = 2.1$), PTBS (poly(*t*-butylstyrene), $k = 2.4$), PS (polystyrene, $k = 2.5$), ActivInk D2200 (polyolefin-polyacrylate, $k = 3.2$) and PMMA (poly(methylmethacrylate), $k = 3.6$). Figure 2a–c and Supplementary Fig. 9 show representative output and transfer current–voltage plots for TFTs fabricated by spin-coating both the semiconductor and the dielectric layers on glass/Au substrates (entries 1–6 in Table 1). Inspection of the output plot at low source–drain voltage, V_{SD} (Fig. 2a, b insets, and Supplementary Fig. 9) reveals excellent electron injection from high-work-function Au contacts with no significant second-order curvatures. This result is typical for all P(NDI2OD-T2)-based TFTs, independent of the gate dielectric. The device contact resistance, estimated from the transfer line analysis method[27], is $\sim$10–100 kΩ cm, which is within those measured for

Table 1 | Performance parameters measured in ambient for P(NDI2OD-T2)-based TGBC TFTs*

Entry	Substrate	Dielectric† (d, nm)	*k*	P(NDI2OD-T2) Deposition (solvent)‡	μ_e§|| ($cm^2 V^{-1} s^{-1}$)	$I_{on}:I_{off}$¶ (log_{10})	V_{on} (V)	V_T (V)	*S* (V per dec.)
1	Glass	CYTOP (450–600)	2.0	Spin-coating (DCB)	0.1–0.25	6–7	5–10	15–20	2–3
2	Glass	PTBS (600–800)	2.4	Spin-coating (DCB)	0.1–0.4	6–7	5–10	15–20	2–3
3	Glass	PS (500–700)	2.5	Spin-coating (DCB)	0.1–0.3	7–8	0–10	10–15	2–3
4	Glass	D2200 (350–500)	3.2	Spin-coating (DCB)	0.2–0.85	6–7	0–5	5–10	1–2
5	Glass	PMMA (600–900)	3.6	Spin-coating (DCB)	0.1–0.25	6–7	0–5	5–10	1–2
6	Glass	PMMA (600–900)	3.6	Spin-coating (Xylene)	0.2–0.45	6–7	−5 to 0	5–10	3–5
7	PET	CYTOP (450–600)	2.0	Spin-coating (DCB)	0.1–0.2	6–7	10–15	20–25	2–3
8	PET	PTBS (600–800)	2.4	Spin-coating (DCB)	0.1–0.5	5–6	10–20	25–35	5–7
9	PET	PS (500–700)	2.5	Spin-coating (DCB)	0.1–0.3	6–7	5–10	15–25	2–3
10	PET	D2200 (350–500)	3.2	Spin-coating (DCB)	0.2–0.5	6–7	0–5	5–10	1–2
11	PET	PMMA (600–900)	3.6	Spin-coating (DCB)	0.1–0.25	6–8	0–5	10–20	2–3
12	PET	D2200 (1,000–1,200)	3.2	Gravure ($DCB\text{-}CHCl_3$)	0.1–0.4	6–7	0–5	5-10	2–3
13	PET	PMMA (600–900)	3.6	Gravure (xylene)	0.1–0.2	5–6	0–5	10–15	2–4
14	PET	D2200 (1,000–1,200)	3.2	Flexo (DCB)	0.1–0.3	5–6	10–20	25–35	8–10
15[e]	PET#	D2200 (1000–1200)	3.2	Inkjet (DCB)	~0.1	~5	~10	~40	~10
16	PET	Gravure D2200 (1,000–1,200)	3.2	Gravure ($DCB\text{-}CHCl_3$)	0.1–0.65	5–7	5–10	30–35	4–6
17	PET☆	Gravure PMMA (700–1,200)	3.6	Gravure (xylene)	0.1–0.2	4–6	0–5	10–15	2–3

* For bottom-gate top-contact TFT performance and fabrication details on conventional Si/SiO_2-OTS substrates ($\mu = 0.01$–$0.08\ cm^2 V^{-1} s^{-1}$), see ref. 37.
† Deposited by spin-coating unless indicated; *d*, the dielectric film thickness.
‡ For solvent ratio see Methods.
§ All electron mobilities (μ_e) and threshold voltages (V_T) are calculated in the saturation regime. The minimum and maximum values are reported.
|| In some instances erratic ambipolar behaviour was recorded although the hole mobilities (μ_h) are $<10^{-4}\ cm^2 V^{-1} s^{-1}$.
¶ Calculated at $V_G = 0.0$–60.0 V and $V_{SD} = 60.0$ V for all devices except those of entry 15 ($V_G = 0.0$–90.0 V and $V_{SD} = 60.0$ V) and 16 ($V_G = 0.0$–80.0 V and $V_{SD} = 60.0$ V).
#From the devices exhibiting acceptable source/drain channel coverage.
☆PEN substrates were also employed for device fabrication.

n-channel perylene-based[27] and oligothiophene-based[39] staggered TFTs using Au contacts. However, contact resistance reduction, and hence improved TFT characteristics, may be possible via alkylthiol treatment of the Au contacts and using low-work-function metals[27,40].

The data collected in Table 1 clearly demonstrate the high performance of P(NDI2OD-T2)-based TGBC TFTs fabricated and measured in ambient conditions. The spin-coated devices based on CYTOP, PS, PTBS, D2200 and PMMA all exhibit very high performance, with average values of $\mu_e \approx 0.2$–$0.5\ cm^2 V^{-1} s^{-1}$, $I_{on}/I_{off} > 10^6$, $V_{on} \approx 0$–15 V, $V_{th} \approx 5$–20 V, and $S < 3$ V per decade with device yields approaching 100%. Several devices from different batches based on PS, PTBS, PMMA and D2200 exhibit μ_e values of 0.45–$0.85\ cm^2 V^{-1} s^{-1}$ with $I_{on}/I_{off} > 10^6$. Note that statistically comparable device performance and yields are achieved for TFTs on PET plastic substrates (entries 7–11 in Table 1, Fig. 2d), which is essential to enable high-performance flexible printed devices (see below). Furthermore, spin-coating or printing the semiconducting polymer from non-chlorinated solvents, such as xylene, also results in excellent TFT characteristics (entries 6 and 13 in Table 1, Supplementary Fig. 10). The μ_e values of these TFTs approach those of the best n-channel devices based on molecular semiconductors vapour-deposited on conventional substrates[26–28], and are only about 2–3 times lower than those of the best single crystal n-channel transistors reported to date ($\mu_e \approx 1.6\ cm^2 V^{-1} s^{-1}$ in vacuum for TCNQ)[41].

An interesting question that arises is what may be the origin of the excellent performance of the P(NDI2OD-T2)-based top-gate devices as compared to bottom-gate top-contact Si/SiO_2-OTS devices ($\mu_e = 0.01$–$0.08\ cm^2 V^{-1} s^{-1}$, Au contacts)[37] and previously investigated perylene-polymer-based bottom-gate top-contact TFTs ($\mu_e = 0.001$–$0.01\ cm^2 V^{-1} s^{-1}$)[33,34]. To attempt to answer this question, we have fabricated top-gate TFTs using as semiconductor the corresponding perylene-bis(dicarboximide)-dithiophene polymer (P(PDI2OD-T2)[37], see chemical structure in Supplementary Fig. 11). TGBC devices based on this polymer exhibit electron mobilities about 20–30 times lower ($\mu_e = 0.01$–$0.02\ cm^2 V^{-1} s^{-1}$) than those of side-by-side fabricated P(NDI2OD-T2)-based TGBC TFTs ($\mu_e = 0.3$–$0.4\ cm^2 V^{-1} s^{-1}$, see Supplementary Fig. 11). Note that P(NDI2OD-T2) Si/SiO_2-OTS bottom-gate devices remain active ~16 weeks after fabrication ($\mu_e \approx 0.01\ cm^2 V^{-1} s^{-1}$), whereas P(PDI2OD-T2) bottom-gate TFTs exhibit an initial $\mu_e \approx 10^{-3}\ cm^2 V^{-1} s^{-1}$ and after 1.5 weeks

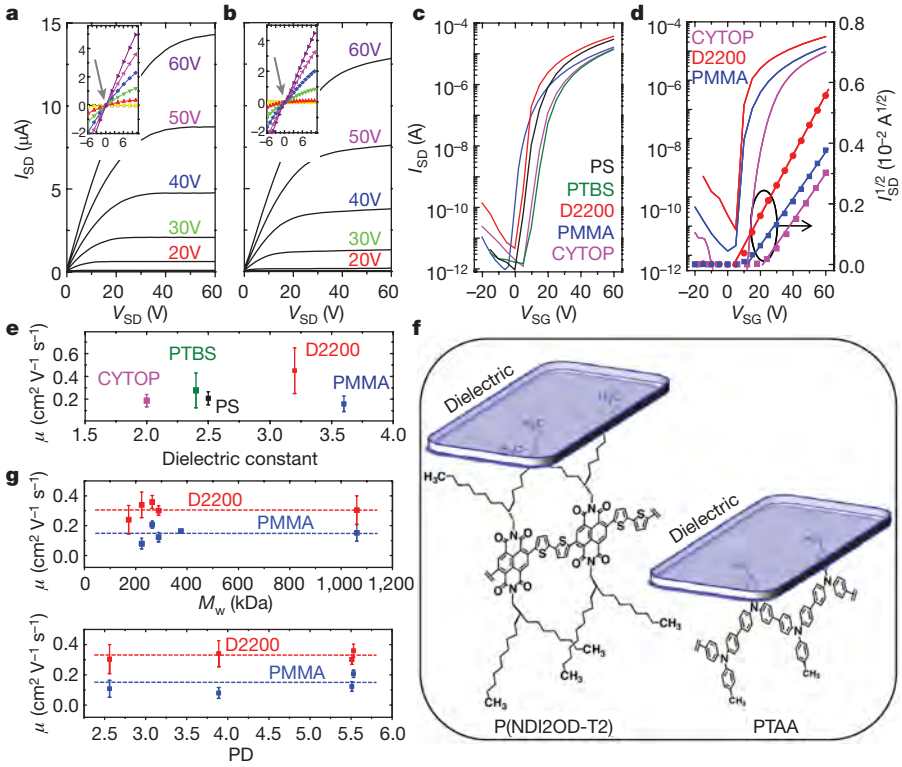


Figure 2 | Performance in ambient conditions of representative TGBC TFT devices with spin-coated P(NDI2OD-T2) semiconductor and various dielectric layers. Channel length (L) and width (W) of these devices are respectively 50 and 1,000 μm. **a**, Current–voltage output plot as a function of V_{SG} for a PMMA-based device on glass. Inset, low-voltage scan (axes as main plot) highlighting the linear I–V characteristics and the line intersection through the axes origin. **b**, Current–voltage output plot as a function of V_{SG} for a PTBS-based device on glass. Inset, scan as **a**. **c**, TFT transfer plots of current versus V_{SG} for representative TGBC device on glass using the indicated polymeric dielectrics. **d**, TFT transfer plot of current versus V_{SG} for representative devices on PET using the indicated polymeric dielectrics. **e**, Plot of the average electron mobility versus the dielectric constant of the indicated dielectrics. Error bars, s.d. **f**, Schematics of P(NDI2OD-T2) and PTAA polymers below the gate dielectric surface. **g**, Plots of the average electron mobility for various P(NDI2OD-T2) batches as a function of the polymer weight-average molecular weight (M_w, top; PD = 2.9–5.5) and polydispersity (PD, bottom; M_w = 220–290 kDa). The broken lines are guides to the eyes. Error bars, s.d.

became mostly inactive[37]. Therefore, we believe that the regioregularity and electronic structure of NDI-based versus PDI-based polymers is clearly at the origin of the overall improved P(NDI2OD-T2) transistor performance. Furthermore, the data of Table 1 demonstrate that top-gate TFTs based on polymeric dielectrics perform better (by about 5–10 times) and are more stable (see below) than bottom-gate TFTs fabricated on Si/SiO_2-OTS substrates, which could be the result of using electron-trapping-free silanol(Si-OH)/carbinol(C-OH) dielectric materials and the different device processing history[19].

TFT performance versus dielectric and polymer architectures

Figure 2e plots the average carrier mobility for several P(NDI2OD-T2) TGBC TFTs versus the dielectric constant of the corresponding gate dielectric material. Interestingly, this plot reveals little sensitivity of μ_e on the k of the gate dielectric, which is a signature of efficient electron transport in this polymer. This result differs substantially from that reported for amorphous triarylamine-based p-channel polymers (PTAA), where the hole mobility decreases by ~20 times when k increases from ~2.0 to ~3.6 (ref. 42). A strong dependence of μ on k has also been observed for rubrene single-crystal TFTs[12]. Variable-temperature current–voltage (I–V) measurements, accurate μ versus V_{SG} analysis, and charge-modulation spectroscopies will shed more light on the details of the charge transport mechanism in this semiconductor[43,44]. However, for PTAA the mobility dependence was explained in terms of broadening of the density of states due to static dipolar disorder in the dielectric[42], whereas for rubrene this was accounted as the result of charge localization and formation of Froehlich polarons caused by the interaction of the charge with the induced dipole moments in the dielectric[12].

A recent theoretical model explains, in a quantitative and comprehensive manner, how the dielectric k and dipoles affect the broadening of the semiconductor density of states and Froehlich polaron energies and incorporates them into a mobility model with excellent agreement with the experimental data for PTAA[43]. Depending on the gate voltage, Froehlich polaron binding energy may contribute to the overall mobility variations. However, this model clearly shows that the impact of both factors in reducing μ when k increases strongly diminish when the distance increases between the site where charge transport occurs (the semiconductor π-conjugated core) and the dielectric surface. It is known that OTFT charge transport is confined in a thin layer (usually ~1 nm) as close as possible to the interface with the dielectric[2]. For amorphous polymers, the chains are randomly distributed without a well-defined orientation with respect to the dielectric surface; yet, for PTAA the conjugated core can approach the dielectric surface very closely whereas for P(NDI2OD-T2) the long branched 2-octyldodecyl substituents separate the NDI-T2 conjugated core from the surface at a distance >1 nm (Fig. 2f, Supplementary Fig. 12). We believe that the significant decoupling of the semiconductor core from the dipoles within the dielectrics is probably at the base of the substantial k-insensitive mobility behaviour.

Given the large electron mobilities achieved using P(NDI2OD-T2) batches with high weight-average molecular weight (M_w) but significantly large polydispersity (PD), an interesting question is whether the TFT performance of this polymer is sensitive to these parameters and whether they can be improved by their further optimization. To our knowledge, no systematic studies have analysed how the polymeric semiconductor PD affects the corresponding TFT properties.

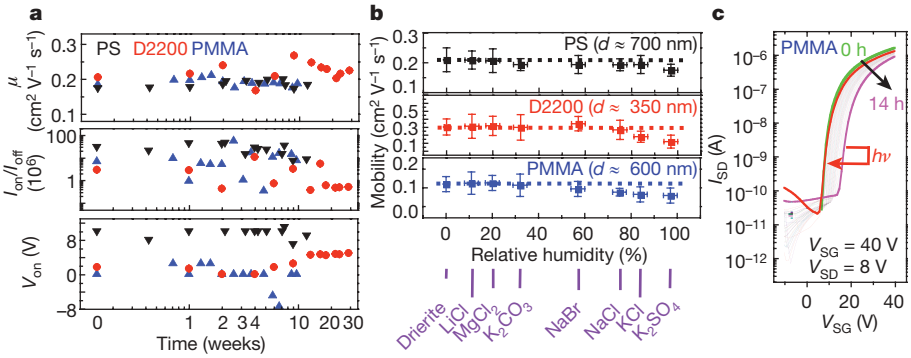


Figure 3 | Stability and bias stress in ambient of representative TGBC TFTs with spin-coated P(NDI2OD-T2) semiconductor and several gate dielectrics. a, Transistor performance parameters versus time plots for three TGBC TFT arrays on glass using PS, D2200 and PMMA as the gate dielectric. Relative humidity RH = 20–60%. The dielectric thicknesses are ~400–800 nm. **b**, Electron mobility versus relative humidity plot for a TGBC TFT array using PS (glass), D2200 (PET) and PMMA (glass) as the gate dielectrics (substrates). *d*, dielectric thickness. The broken lines are guides to the eyes. Error bars, s.d. The chamber RH, in which the devices were stored for 24 h before testing, was controlled by using aqueous saturated solutions of the indicated inorganic salts. Temperature is 22–26 °C. **c**, Transfer plots (L = 50 μm, W = 1,000 μm) for a TGBC TFT using PMMA as the gate dielectric with the indicated stress conditions. RH ≈ 55%. Green line (before stress), magenta line (after 14 h stress), red line (after light irradiation).

Furthermore, no details are available regarding the M_w dependence of the mobility for amorphous p-channel polymers. On the other hand, fundamental studies have been carried out for semicrystalline polymers such as P3HT[44,45], where the carrier mobility for TGBC TFTs increases by >100 times when going from low- to high-M_w polymer samples and saturates at ~0.1 $cm^2 V^{-1} s^{-1}$ for M_w = 52–270 kDa (film processed from trichlorobenzene)[44]. To answer this question, we have synthesized several P(NDI2OD-T2) batches exhibiting similar M_w (~220–290 kDa) and different PD values (2.6–5.6) as well as batches with M_w varying from ~170 kDa to >1 MDa (see Supplementary Information). Figure 2g shows the average μ_e for TGBC TFTs based on these P(NDI2OD-T2) batches and using PMMA and D2200 as the gate dielectrics. For both dielectrics, very little TFT performance variations with M_w and PD are observed; this is in agreement with the amorphous nature of both the high- and low-M_w polymer films on Au/glass substrates (see XRD in Supplementary Fig. 13). However, the insensitivity of the device performance on the polymer chain length extension (for M_w > 170 kDa) and distribution is of extreme importance for facile polymer large-scale synthesis and batch to batch reproducibility of the TFT characteristics.

TFT stability data in ambient conditions

A relevant issue for organic semiconductor-based electronics, but particularly challenging for n-channel transistors, is stability. Although bias stress effects are certainly of paramount importance for both p- and n-channel TFTs[46–48], device shelf stability is a prerequisite for inexpensive device processing and encapsulation in ambient conditions[18]. For this new semiconductor, a set of PS-, D2200- and PMMA-based TGBC TFTs on glass (entries 3–5 in Table 1) was stored in ambient conditions and the device performance was monitored periodically over several weeks after fabrication. Figure 3a plots the average TFT performance parameters (μ_e, I_{on}/I_{off} and V_{on}) for these arrays over 9 (PMMA), 12 (PS) and 28 (D2200) week time periods, remarkably showing no appreciable statistical variations of μ_e and I_{on}/I_{off} and, in the case of D2200-based TFTs, a small V_{on} increase after ~10 weeks (2.5 V → 4.5 V). These data demonstrate large mobility and stable n-channel TFT characteristics independent of the dielectric material. Furthermore, the stability of these devices in air was monitored under atmospheres of increased relative humidity (RH, ~0% → 98%). Figure 3b plots the electron mobility versus RH for a set of devices based on the same dielectrics on glass/PET (entries 3 and 5/10 in Table 1). These plots show no substantial erosion of the device characteristics up to ~70% RH, demonstrating that these TGBC TFTs operate properly in ambient conditions in the presence of dioxygen and considerable H_2O concentrations. At ~98% RH the electron mobilities drop by ~2–3 times, which is consistent with changes in the dielectric thickness (see *d* values in Fig. 3b). We believe that the appropriate polymer electronic structure, uniform film morphologies, and the use of proper dielectric materials, combined with the self-encapsulated top-gate TFT architecture, are at the base of the excellent device stability.

We have also investigated the bias stress of P(NDI2OD-T2)-based TFTs in ambient conditions and RH of ~55% using PMMA (the highest *k* polymer in this series) as the gate dielectric (Fig. 3c). When applying a constant V_{SG} of 40 V and a V_{SD} of 8 V, after 3 and 14 h stress, the I_{on} current decreases from 1.59 μA (V_{on} = 7 V) to 1.39 μA (−13%, V_{on} = 9 V) and 0.90 μA (−43%, V_{on} = 14 V), respectively. Although bias stress is observed, it is not more severe than recently reported for other p-channel polymers, such as polythiophenes and F8T2[46,47]. Furthermore, by simply irradiating the sample with microscope white light for 10 min, the device performance mostly recovers (I_{on} = 1.30 μA and V_{th} = 7 V), indicating that the stress is not associated with significant decomposition of the polymer radical anion at the semiconductor-dielectric interface but rather with charge trapping, probably within the semiconductor film[48].

Printed TFTs and complementary invertors

As P(NDI2OD-T2) exhibits a unique combination of large solubility and comparable carrier mobilities on rigid and flexible substrates, various printing methods to process the semiconductor film were explored for TFT fabrication. Note that different printing techniques require substantially different ink rheological properties, ranging from very viscous formulations used in flexography to less viscous gravure-printing inks to far more dilute solutions suitable for inkjet printing[18]. Therefore, it cannot be assumed a priori that a polymer functioning well in spin-coated devices, and hence solution-processable, will necessarily be printable. Because of our previous experience in gravure printing dielectric and semiconductor formulations, this technique was investigated first. For successful TFT fabrication by printing, it is necessary to achieve a smooth and uniform semiconductor film morphology so that a pinhole-free gate dielectric film can be deposited on top. Figure 4a (panels 1 to 6) provides an example of gravure-printing optimization, demonstrating the importance of the semiconductor ink formulation viscosity and gravure cylinder cell volume to afford proper polymer film morphologies. Figure 4b and Supplementary Fig. 14 show the AFM images of optimized gravure-printed versus spin-coated P(NDI2OD-T2) films on PET/Au substrates. Both films exhibit a fibre-like morphology similar to that observed on conventional Si-SiO_2 substrates. Furthermore, the gravure-printed films exhibit extensive pitting (depth only ~2 nm for ~80-nm-thick film), possibly resulting from

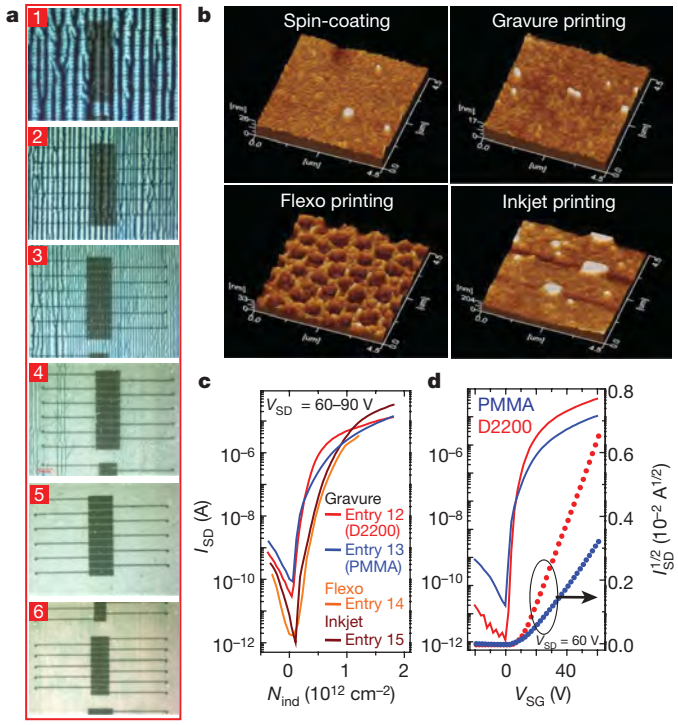


Figure 4 | P(NDI2OD-T2) film morphologies and TGBC TFT performance for polymer films/devices fabricated using various solution-processing techniques on PET/Au substrates. ($L = 50$ μm, $W = 1{,}000$ μm.) **a**, Optical images of P(NDI2OD-T2) films with underlying Au source–drain contacts gravure-printed from a 2% w/w polymer DCB-$CHCl_3$ 50-50 v/v solution and using printing disks with decreasing gravure cell depth/volume ratio (depth D in μm, volume V in ml m^{-2}). Image 1, $D/V = 80/23.3$; image 2, 65/15.6; image 3, 45/6.1; image 4, 40/8.8; image 5, 30/4.1; image 6, 10/0.5. The optimized film morphology corresponds to image 5. **b**, Atomic force microscopy (AFM) images of spin-coated and printed films ('flexo' indicates 'flexographic'). **c**, Representative TFT transfer plots of current versus carrier density (N_{ind}) of various gravure-, flexo- and inkjet-printed TGBC devices. **d**, TFT transfer plot of current versus V_{SG} for representative TGBC TFTs with gravure-printed semiconductor and dielectric layers.

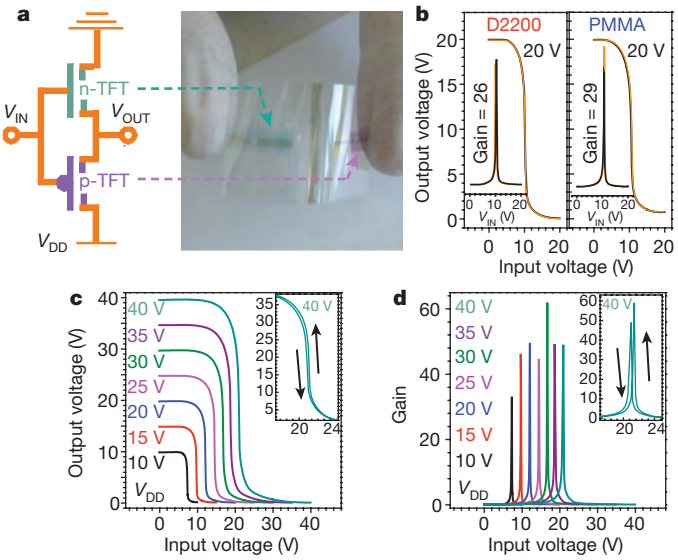


Figure 5 | P3HT (p-channel)-P(NDI2OD-T2) (n-channel) complementary inverters. **a**, Schematic electrical connections of the inverters, and optical image of the gravure-printed device before common gate deposition. **b**, Static switching characteristics of the inverter where both the semiconductors and the indicated dielectrics are gravure-printed (insets, gain of the corresponding device). Red (forward) and black (reverse) scans. **c**, Static switching characteristics of a spin-coated inverter based on PMMA. **d**, Gains of the corresponding spin-coated devices. Insets in **c**, **d**, are expanded views with the same axes as the main plots.

the gravure cylinder topography. However, as both films are comparably smooth (r.m.s. roughness ~1–2 nm for a 20×20 μm^2 area), TGBC TFTs with a ~1-μm-thick PMMA or D2200 dielectric layer can be fabricated with high fidelity (μ_e values ~0.1–0.2 cm^2 V^{-1} s^{-1}, entries 12 and 13 in Table 1).

We have also demonstrated TGBC TFTs by flexographic and inkjet printing the semiconductor layer (entries 14 and 15 in Table 1). The flexographically-printed film morphology exhibits in relief the circular patterns of the flexographic printing plate surface (Fig. 3b and Supplementary Fig. 13). Note that while flexographically-printed P(NDI2OD-T2) films are quite uniform and only slightly less smooth than the spin-coated/gravure-printed films (r.m.s. roughness of 4–6 nm, Fig. 4b), inkjet-printing with our instrumentation results in rougher and far less uniform morphologies with only partial uniform coverage of the channel region (r.m.s. roughness of 8–9 nm, Fig. 4b). Both flexographically- and inkjet-printed TFTs afford electron mobilities >0.1 cm^2 V^{-1} s^{-1} and acceptable device characteristics (Fig. 4c). The excellent performance of the TFTs based on gravure-deposited semiconductor layers led us to explore device fabrications where both the semiconductor and the dielectric layers are gravure-printed (entries 16 and 17 in Table 1). The transfer plots of representative devices based on PMMA and D2200 (Fig. 4d and Supplementary Fig. 15) demonstrate sharp turn-on, good saturation, negligible I–V hysteresis (<5%), and excellent behaviour at acceptable biases ($\mu_e = 0.1$–0.65 cm^2 V^{-1} s^{-1} and $I_{on}/I_{off} > 10^6$), confirming the good interface quality between the two gravure-printed films.

Finally, p-channel TGBC polymeric transistors were fabricated using poly(3-hexylthiophene) (P3HT) as the hole transporting semiconductor and Au/PMMA and Au/D2200 as the contact/dielectric materials. These TFTs exhibit hole mobilities of ~0.02–0.1 cm^2 V^{-1} s^{-1}, $I_{on}/I_{off} = 10^2$–10^3 and V_{on} of −10 to −5 V (Supplementary Fig. 16). Owing to our n-channel polymer robustness, excellent TFT performance with high-work-function metal contacts, and compatibility of PMMA and D2200 with both p- and n-channel semiconductors, polymeric complementary logic with a single materials set and operating in ambient conditions is possible for the first time. To demonstrate materials and processing generality, complementary inverters having P3HT (p-channel, BASF Sepiolid P 100) and P(NDI2OD-T2) (n-channel, Polyera ActivInk N2200) transistors were fabricated by (1) spin-coating both semiconductors and the dielectric layer (PMMA and D2200), and (2) gravure printing all materials (except the contacts). Figure 5a shows an optical image of a printed inverter without the common top-gate layer used as input voltage (V_{IN}). For both devices, inverter response is clearly observed for switching between logic '1' (10–40 V) and logic '0' (0 V) (Fig. 5b, c and Supplementary Fig. 17). All inverters show remarkably small hysteresis, reflecting the transistor threshold voltage stability. For both dielectrics, the voltage gains for the gravure-printed (Fig. 4b insets) and spin-coated (Fig. 5d and Supplementary Fig. 16) devices are very large ($dV_{OUT}/dV_{IN(max)} > 25$ and 60, respectively), implying that these devices could be used to switch subsequent stages in more complex logic circuits.

Overview

In conclusion, we have demonstrated top-gate n-channel polymeric TFTs exhibiting field-effect mobilities, processability, compatibility with various top-gate dielectric materials, and operational/stress stability approaching or even surpassing those based on the best p-channel polymers. Top-gate bottom-contact transistors exhibiting electron mobilities >0.1 cm^2 V^{-1} s^{-1} (up to ~0.85 cm^2 V^{-1} s^{-1}) in ambient conditions with $I_{on}/I_{off} > 10^6$ have been fabricated by spin-coating as well as gravure, flexographic and inkjet printing the semiconducting layer, demonstrating great processing versatility. The

TFT performance was monitored in ambient conditions and under different relative humidities, demonstrating remarkable stability. Finally, the first spin-coated and gravure-printed polymeric semiconductor complementary inverters exhibiting large gains (>25–60) and operating in ambient conditions have been realized. We believe that the discovery of this n-channel material has answered several questions about field-effect electron transport capabilities of polymeric semiconductors and that new fundamental studies will be possible by high-level computational modelling[49], employing unique device structures, and elegant spectroscopies[50]. Furthermore, we are convinced that the combination of this material with previously developed high-performance p-channel polymers[23–25] and optimized device architectures[14–17] will open unprecedented opportunities for printed electronics.

Note added in proof: The synthesis of a naphthalenedicarboximide-dithiophene (NDIR-T2)-based polymer has recently been reported[51].

METHODS SUMMARY

The polymer synthesis, film processing and transistor device fabrication details are reported in Methods, Supplementary Information, and ref. 37. The field-effect mobility in saturation ($V_{SG} - V_T < V_{SD}$) is calculated from equation (1):

$$\left.\frac{\partial\sqrt{I_{SD}}}{\partial V_{SG}}\right|_{V_{SD}} = \sqrt{\frac{W}{2L}C_i\mu(\mathrm{sat})} \quad (1)$$

where I_{SD} is the source–drain current, L (here 25–75 μm) and W (here 0.5–1.5 mm) are respectively the TFT channel length and width, V_T is the threshold voltage, and C_i the dielectric capacitance per unit area. The threshold voltage V_T can be calculated from the x-axis intercept of the square root of I_{SD} versus V_{SG} line. The field-effect mobility in the linear region ($V_G - V_T \gg V_D$) is calculated using the device transconductance (g_m) at a specific V_{SD} from equation (2):

$$g_m = \left.\frac{\partial I_{SD}}{\partial V_{SG}}\right|_{V_{SD}} = \frac{W}{L}C_i\mu(\mathrm{lin})V_{SD} \quad (2)$$

The turn-on voltage (V_{on}) is determined from the logarithmic plot of the subthreshold drain current versus V_g extrapolated at the x-axis intercept. The device on/off current ratio, I_{on}/I_{off}, is typically reported as 10^x and is calculated from the ratio between the I_{SD} at 60 V (unless indicated) and the I_{SD} at 0 V. The sub-threshold swing S is calculated from equation (3):

$$S = \frac{\mathrm{d}V_{SG}}{\mathrm{d}(\log I_{SD})} \quad (3)$$

and it measures of how rapidly the device switches from the off state to the on state in the region of exponential current increase and is typically reported in V per decade or mV per decade.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 31 July; accepted 12 December 2008.
Published online 21 January 2009.

1. Malliaras, G. & Friend, R. H. An organic electronics primer. *Phys. Today* **58,** 53–58 (2005).
2. Klauk, H. *Organic Electronics: Materials, Manufacturing and Applications* (Wiley-VCH, 2006).
3. Sirringhaus, H. *et al.* Two-dimensional charge transport in self-organized, high-mobility conjugated polymers. *Nature* **401,** 685–688 (1999).
4. Briseno, A. L. *et al.* Patterning organic single-crystal transistor arrays. *Nature* **444,** 913–917 (2006).
5. Vikram, C. S. *et al.* Elastomeric transistor stamps: Reversible probing of charge transport in organic crystals. *Science* **303,** 1644–1646 (2004).
6. Muccini, M. A bright future for organic field-effect transistors. *Nature Mater.* **5,** 605–613 (2006).
7. Zaumseil, J., Friend, R. H. & Sirringhaus, H. Spatial control of the recombination zone in an ambipolar light-emitting organic transistor. *Nature Mater.* **5,** 69–74 (2006).
8. Kim, C., Facchetti, A. & Marks, T. J. Polymer gate dielectric surface viscoelasticity modulates pentacene transistor performance. *Science* **318,** 76–80 (2007).
9. Gundlach, D. J. *et al.* Contact-induced crystallinity for high-performance soluble acene-based transistors and circuits. *Nature Mater.* **7,** 216–221 (2008).
10. Dimitrakopoulos, C. D. *et al.* Low-voltage organic transistors on plastic comprising high-dielectric constant gate insulators. *Science* **283,** 822–824 (1999).
11. Dodabalapur, A., Katz, H. E., Torsi, L. & Haddon, R. C. Organic heterostructure field-effect transistors. *Science* **269,** 1560–1562 (1995).
12. Hulea, I. N. *et al.* Tunable Froehlich polarons in organic single-crystal transistors. *Nature Mater.* **5,** 982–986 (2006).
13. See, K. C., Becknell, A., Miragliotta, J. & Katz, H. E. Enhanced response of n-channel naphthalenetetracarboxylic diimide transistors to dimethyl methylphosphonate using phenolic receptors. *Adv. Mater.* **19,** 3322–3327 (2007).
14. Crone, B. *et al.* Large-scale complementary integrated circuits based on organic transistors. *Nature* **403,** 521–523 (2000).
15. Gelinck, G. H. *et al.* Flexible active-matrix displays and shift registers based on solution-processed organic transistors. *Nature Mater.* **3,** 106–109 (2004).
16. Bao, Z., Rogers, J. A. & Katz, H. E. Printable organic and polymeric semiconducting materials and devices. *J. Mater. Chem.* **9,** 1895–1904 (1999).
17. Rogers, J. A. *et al.* Paper-like electronic displays: Large-area rubber-stamped plastic sheets of electronics and microencapsulated electrophoretic inks. *Proc. Natl Acad. Sci. USA* **98,** 4835–4840 (2001).
18. Gamota, D. R. Brazis, P., Kalyanasundaram, X. & Zhang, J. (eds) *Printed Organic and Molecular Electronics* (Kluwer Academic, 2004).
19. Garnier, F., Hajlaoui, R., Yassar, A. & Srivastava, P. All-polymer field-effect transistor realized by printing techniques. *Science* **265,** 1864–1866 (1994).
20. Sivaramakrishnan, S. *et al.* Controlled insulator-to-metal transformation in printable polymer composites with nanometal clusters. *Nature Mater.* **6,** 149–155 (2007).
21. Facchetti, A., Yoon, M.-H. & Marks, T. J. Gate dielectrics for organic field-effect transistors: New opportunities for organic electronics. *Adv. Mater.* **17,** 1705–1725 (2005).
22. Cho, J. *et al.* High-capacitance ion gel gate dielectrics with faster polarization response times for organic thin film transistors. *Adv. Mater.* **20,** 686–690 (2008).
23. McCulloch, I. *et al.* Liquid-crystalline semiconducting polymers with high charge-carrier mobility. *Nature Mater.* **5,** 328–333 (2006).
24. Pan, H. *et al.* Low-temperature, solution-processed, high-mobility polymer semiconductors for thin-film transistors. *J. Am. Chem. Soc.* **129,** 4112–4113 (2007).
25. Dhoot, A. S. *et al.* Beyond the metal-insulator transition in polymer electrolyte gated polymer field-effect transistors. *Proc. Natl Acad. Sci. USA* **103,** 11834–11837 (2006).
26. Ando, S. *et al.* n-Type organic field-effect transistors with very high electron mobility based on thiazole oligomers with trifluoromethylphenyl groups. *J. Am. Chem. Soc.* **127,** 14996–14997 (2005).
27. Chesterfield, R. J. *et al.* Organic thin film transistors based on n-alkyl perylene diimides: Charge transport kinetics as a function of gate voltage and temperature. *J. Phys. Chem. B* **108,** 19281–19292 (2004).
28. Anthopoulos, T. D. *et al.* High performance n-channel organic field-effect transistors and ring oscillators based on C_{60} fullerene films. *Appl. Phys. Lett.* **89,** 213504 (2006).
29. Waldauf, C., Schilinsky, P., Perisutti, M., Hauch, J. & Brabec, C. J. Solution-processed organic n-type thin-film transistors. *Adv. Mater.* **15,** 2084–2088 (2003).
30. Newman, C. R. *et al.* Introduction to organic thin film transistors and design of n-channel organic semiconductors. *Chem. Mater.* **16,** 4436–4451 (2004).
31. Chua, L.-L. *et al.* General observation of n-type field-effect behavior in organic semiconductors. *Nature* **434,** 194–199 (2005).
32. Letizia, J. n-Channel polymers by design: Optimizing the interplay of solubilizing substituents, crystal packing, and field-effect transistor characteristics in polymeric bithiophene-imide semiconductors. *J. Am. Chem. Soc.* 130, 9679–9694 (2008).
33. Zhan, X. *et al.* A high-mobility electron-transport polymer with broad absorption and its use in field-effect transistors and all-polymer solar cells. *J. Am. Chem. Soc.* **129,** 7246–7247 (2007).
34. Huttner, S., Sommer, M. & Thelakkat, M. n-type organic field effect transistors from perylene bisimide block copolymers and homopolymers. *Appl. Phys. Lett.* **92,** 093302 (2008).
35. Briseno, A. *et al.* Self-assembly, molecular packing, and electron transport in n-type polymer semiconductor nanobelts. *Chem. Mater.* **20,** 4712–4719 (2008).
36. Klauk, H., Zschieschang, U., Pflaum, J. & Halik, M. Ultralow-power organic complementary circuits. *Nature* **445,** 745–748 (2007).
37. Chen, Z., Zheng, Y., Yan, H. & Facchetti, A. Naphthalenedicarboximide- vs. perylenedicarboximide-based copolymers. Synthesis and semiconducting properties in bottom-gate n-channel organic transistors. *J. Am. Chem. Soc.* **131,** 8–9 (2009).
38. Jones, B. A., Facchetti, A., Wasielewski, M. R. & Marks, T. J. Tuning orbital energetics in arylene diimide semiconductors. Materials design for ambient stability of n-type charge transport. *J. Am. Chem. Soc.* **129,** 15259–15278 (2007).
39. Dholakia, G. R., Meyyappan, M., Facchetti, A. & Marks, T. J. Monolayer to multilayer nanostructural growth transition in n-type oligothiophenes on Au(111) and implications for organic field-effect transistor performance. *Nano Lett.* **6,** 2447–2455 (2006).
40. Stoliar, P. *et al.* Charge injection across self-assembly monolayers in organic field-effect transistors: Odd-even effects. *J. Am. Chem. Soc.* **129,** 6477–6484 (2007).
41. Menard, E. *et al.* High-performance n- and p-type single-crystal organic transistors with free-space gate dielectrics. *Adv. Mater.* **16,** 2097–2101 (2004).

42. Veres, J. *et al.* Low-k insulators as the choice of dielectrics in organic field-effect transistors. *Adv. Funct. Mater.* **13,** 199–204 (2003).
43. Richards, T., Bird, M. & Sirringhaus, H. A quantitative analytical model for static dipolar disorder broadening of the density of states at organic heterointerfaces. *J. Chem. Phys.* **128,** 234905 (2008).
44. Chang, J.-F., Sirringhaus, H., Giles, M., Heeney, M. & McCulloch, I. Relative importance of polaron activation and disorder on charge transport in high-mobility conjugated polymer field-effect transistors. *Phys. Rev. B* **76,** 205204 (2007).
45. Kline, R. J. *et al.* Dependence of regioregular poly(3-hexylthiophene) film morphology and field-effect mobility on molecular weight. *Macromolecules* **38,** 3312–3319 (2005).
46. Tse, N. N. *et al.* Gate bias stress effects due to polymer gate dielectrics in organic thin-film transistors. *J. Appl. Phys.* **103,** 044506 (2008).
47. Richards, T. & Sirringhaus, H. Bias-stress induced contact and channel degradation in staggered and coplanar organic field-effect transistors. *Appl. Phys. Lett.* **92,** 023512 (2008).
48. Salleo, A. & Street, R. A. Light-induced bias stress reversal in polyfluorene thin-film transistors. *J. Appl. Phys.* **94,** 471–479 (2003).
49. Bredas, J.-L. *et al.* Charge transport in organic semiconductors. *Chem. Rev.* **107,** 926–952 (2007).
50. Westenhoff, S., Howard, I. A. & Friend, R. H. Probing the morphology and energy landscape of blends of conjugated polymers with sub-10 nm resolution. *Phys. Rev. Lett.* **101,** 016102 (2008).
51. Guo, X. & Watson, M. D. Conjugated polymers from naphthalene bisimide. *Org. Lett.* **10,** 5333–5336 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank P. Inagaki for his leadership, T. J. Marks for discussions and P. Eckerle and BASF Future Business for their support.

Author Contributions H.Y. supervised device fabrication and analysis, performed the humidity tests, and fabricated the complementary inverters. Z.C. designed and synthesized the semiconductor polymer. Y.Z. fabricated the devices by spin-coating and monitored the stability in ambient conditions. C.N. fabricated most of the gravure- and inkjet-printed devices and acquired all AFM images. J.Q. optimized NDI monomer and dielectric synthesis. F.D. and M.K. supported the synthetic efforts. A.F. directed the project and wrote the manuscript.

Author Information. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.F. (afacchetti@polyera.com).

METHODS

Synthesis, batch 1 of P(NDI2OD-T2). Chemical name, poly{[*N*,*N*′-bis(2-octyldodecyl)-1,4,5,8-naphthalene-bis(dicarboximide)-2,6-diyl]-alt-5,5′-(2,2′-bithiophene)}. Under argon, a mixture of NDI2OD-Br_2 (76 mg, 0.077 mmol, pure by elemental analysis)[37], 5,5′-bis(trimethylstannyl)-2,2′-bithiophene (38 mg, 0.077 mmol, pure by elemental analysis), and $Pd(PPh_3)_2Cl_2$ (2 mg, 0.003 mmol) in anhydrous toluene (4 ml) was stirred at 90 °C for 2 days. Bromobenzene (0.3 ml) was then added and the reaction mixture was maintained at 90 °C for an additional 12 h. Upon cooling to room temperature, a solution of potassium fluoride (1 g) in water (2 ml) was added and the mixture stirred at room temperature for one hour before it was diluted with chloroform (100 ml). The resulting solution was washed with water (80 ml × 3), dried over anhydrous sodium sulphate, and concentrated on a rotary evaporator. The residue was dissolved in chloroform (20 ml) and the resulting solution was precipitated in methanol (50 ml). The obtained blue solid product was further purified by Soxhlet extraction with acetone for 48 h. The isolated solid material was dissolved in chloroform (40 ml) and the resulting mixture was heated to boiling point. Upon cooling to room temperature, this chloroform solution was filtered through a syringe filter (5 μm), and the filtrate was precipitated in methanol (40 ml). The precipitate was collected by filtration, washed with methanol, and dried in vacuum, leading to a deep blue solid as the product (67 mg, yield 88%). ^{1}H NMR ($C_2D_2Cl_4$, 500 MHz): δ; 8.53–8.84 (m, br, 2H), 7.20–7.48 (br, 4H), 4.13 (s, br, 2H), 2.00 (s, br, 4H), 1.05–1.30 (s, br, 64H), 0.87 (s, br, 12H). GPC: M_n = 52.5 kDa, M_w = 288.9 kDa, PDI = 5.51. Elemental analysis (calc. C, 75.26; H, 8.96; N, 2.83): found C, 75.19; H, 9.12; N, 2.81.

Materials. Dichlorobenzene (DCB), chloroform ($CHCl_3$), chlorobenzene (CB), xylene, and acetates used for dielectric and semiconductor formulations were purchased from Sigma Aldrich and distilled before use. Solutions of P(NDI2OD-T2) were prepared by Polyera Corporation (available under the trade name ActivInk N2200, see below) and solutions of P3HT were prepared by BASF (available under the trade name Sepiolid P 100, see below). The polymeric dielectrics were obtained from: CYTOP (product no. CTL-809M from Asahi Glass), PTBS (poly(4-tert-butyl-styrene), product no. 369705 from Aldrich, M_w = 50–100 kDa) PS (polystyrene, product no. 331651 from Aldrich, M_w = 35kDa), ActivInk D2200 (Polyera Corporation), PMMA (poly(methyl methacrylate), product no. 182230 from Aldrich, M_w = 120 kDa) and the solutions, filtered through a 0.2 μm size syringe filter before use, were prepared by Polyera Corporation.

TFT fabrication. All processes, except metal evaporation and drying steps, were performed in ambient conditions in a conventional chemistry hood. Top-gate bottom-contact TFTs were fabricated on glass (PGO glass or other sources) and PET (DuPont or other sources). Depending on the substrate source, they were used as received or first planarized by spin-coating a ~400-nm-thick Polyera ActivInk 1100 film (solution concentration ~80–110 mg ml^{-1} in dioxane, spin rate = 1,500–2,000 r.p.m., photocured at λ = 300 nm, dried at 110 °C for 10 min) followed by thermally evaporated Au source–drain contacts (30 nm thick). Channel lengths and widths are 25–75 μm and 0.5–1.5 mm, respectively, to afford W/L = 20. These substrates were coated with the semiconductor layer deposited by spin-coating (concentration ~10–20 mg ml^{-1} in DCB, 1,500–2,000 r.p.m.; DCB-CB 90:10 v/v mixture could also be used depending on the PET substrate source/batch), gravure printing (concentration ~1–2% w/w in DCB-$CHCl_3$ 50:50 v/v mixture, anilox force = 50–100 N, printing speed 0.2 m s^{-1}, anilox cylinder 402.110 IGT printer), flexo printing (concentration ~3–5% w/w in DCB, anilox force = 100–150 N, printing force = 30–100 N, printing speed 0.2 m s^{-1}, anilox cylinder 402.110 IGT printer), and inkjet printing (concentration ~0.1–0.2% w/w in DCB, droplet size = 5 pl, Dimatix 2800 series printer). Typical semiconductor film thicknesses are 40–120 nm. Next, the dielectric layer was spin-coated using the following conditions: CYTOP (commercially available formulation, CTL-809M, 5,000–7,000 r.p.m.), PS (concentration ~50–80 mg ml^{-1} in EtOAc or PrOAc, 1,000–2,000 r.p.m.), PTBS (concentration ~50–80 mg ml^{-1} in PrOAc, 1,000–2,000 r.p.m.), ActivInk D2200 (concentration ~20–60 mg ml^{-1} in hydrocarbon-acetate formulations, 1,500–2,000 r.p.m.), PMMA (concentration ~60–80 mg ml^{-1} in EtOAc or PrOAc, 1,000–2,000 r.p.m.). For the gravure printing experiments PMMA (concentration ~50–100 mg ml^{-1} in EtOAc) or ActivInk D2200 (concentrations ~40–100 mg ml^{-1} in hydrocarbon-acetate formulations) were printed using the following conditions: anilox force = 70–100 N, printing speed = 0.2–0.8 m s^{-1}) to afford 700–1,200 nm thick films. The semiconductor and dielectric films were dried at 110 °C (60 °C for PS-based devices) overnight in a vacuum oven (<5 mtorr). The film dielectric constant (k) of the gate dielectric is shown in Table 1. The device structure was completed by vapour deposition of patterned Au gate contacts (~30 nm thick) through a shadow mask.

The P3HT (p-channel) transistors were fabricated by spin-coating (concentration ~1% w/w in DCB, 1,500–2,000 r.p.m.) and gravure printing (concentration ~1–2% w/w in DCB). The n-channel devices were fabricated as described above. Spin-coated inverters were fabricated by covering a selected substrate area during the deposition of the other semiconductor, whereas gravure printed inverters were fabricated by printing each semiconductor separately on different substrate areas. The resulting p- and n-channel TFTs were connected with a common vapour-deposited gate line (Au, 30 nm).

Film and device characterization. A Keithley 4200 semiconductor characterization system was used to perform all electrical/TFT characterizations of the top gate devices. The capacitance of the dielectric film was measured using a GLK Model 3000 digital capacitance meter. The 4200 SCS system consists of three source measurement units, all of which are supplied with remote pre-amplifiers. The other major component of the test system is a Signatone probe station. Triax cable and probes were used for all electrodes to provide the first level of shielding. A dark/metal box enclosure was used to avoid light exposure and to further reduce environmental noise. The dark box had a triax cable feedthrough panel to maintain consistent triax shielding all the way from the preamps to the end of triax probe tips. Thin film XRD characterization was performed using a Rigaku ATXG thin film diffractometer with Ni-filtered Cu Kα radiation. AFM images were taken from a JEOL-SPM5200 with a Si cantilever. Film thickness was determined by profilometry using a Veeco Dektak 150.

ARTICLES

The unfolded protein response signals through high-order assembly of Ire1

Alexei V. Korennykh[1,3], Pascal F. Egea[1], Andrei A. Korostelev[4], Janet Finer-Moore[1], Chao Zhang[2,3], Kevan M. Shokat[2,3], Robert M. Stroud[1] & Peter Walter[1,3]

Aberrant folding of proteins in the endoplasmic reticulum activates the bifunctional transmembrane kinase/endoribonuclease Ire1. Ire1 excises an intron from *HAC1* messenger RNA in yeasts and *Xbp1* messenger RNA in metozoans encoding homologous transcription factors. This non-conventional mRNA splicing event initiates the unfolded protein response, a transcriptional program that relieves the endoplasmic reticulum stress. Here we show that oligomerization is central to Ire1 function and is an intrinsic attribute of its cytosolic domains. We obtained the 3.2-Å crystal structure of the oligomer of the Ire1 cytosolic domains in complex with a kinase inhibitor that acts as a potent activator of the Ire1 RNase. The structure reveals a rod-shaped assembly that has no known precedence among kinases. This assembly positions the kinase domain for *trans*-autophosphorylation, orders the RNase domain, and creates an interaction surface for binding of the mRNA substrate. Activation of Ire1 through oligomerization expands the mechanistic repertoire of kinase-based signalling receptors.

Approximately one-third of all proteins in eukaryotes enter the endoplasmic reticulum (ER) for processing and folding. The quality of protein folding is monitored by the ER-membrane-resident kinase/RNase Ire1, which is activated by misfolded proteins. On activation, Ire1 initiates a non-spliceosomal mRNA splicing reaction. Translation of the spliced mRNA yields an unfolded protein response (UPR)-specific transcription factor, termed Hac1 (ref. 1) in yeasts and Xbp1 (ref. 2) in metazoans, that induces a comprehensive transcriptional program to adjust the protein-folding capacity of the ER according to need. The UPR is activated in cancers[3,4], viral infections[5], protein-folding diseases[6,7] and other cellular anomalies[8,9].

Under ER stress conditions, the ER-lumenal domain of Ire1 acts as a sensor of unfolded proteins[10] (Fig. 1a). It crystallizes as a polymer that has two distinct crystallographic interfaces important for function[10]. This feature can explain an early observation of oligomerization of Ire1 during the UPR[11] and provide a structural rationalization of Ire1 organization into UPR-induced clusters (foci) that can be observed by live cell imaging[12,13]. Oligomerization of the ER-lumenal domain of Ire1 has been proposed to promote dimerization and activation of the kinase/RNase domains[10,14] analogous to other cell-surface signalling receptors[15]. Additionally, *trans*-autophosphorylation and binding of ADP are thought to contribute to activation of the Ire1 RNase[16,17].

A recent crystal structure of the kinase/RNase domain of Ire1 reveals a two-fold symmetric dimer with a back-to-back arrangement of the kinase domains, compactly attached to an RNase dimer proposed to have two independent active sites[14]. The back-to-back arrangement of the kinases in the dimer is unexpected because it positions the phosphorylation sites in the activation loops more than 40 ångström (Å) away from the active site of the partnering molecule in the dimer. This arrangement does not seem to be productive for the *trans*-autophosphorylation of Ire1 observed *in vivo*[11]. We propose that a different Ire1 dimer enables the *trans*-autophosphorylation reaction (below). Dimerization of the RNase domains has been proposed to match functionally the conserved pair of splice sites in *HAC1*/*Xbp1* mRNA[14] (Fig. 1b). Our results indicate an alternative explanation because we observe fully reactive RNA substrates that contain only a single splice site, as well as poorly reactive RNA substrates that contain dual splice sites. In this work, we combine several approaches to show that the cytosolic region of Ire1 from *Saccharomyces cerevisiae* undergoes spontaneous oligomerization that activates Ire1 for signalling in the UPR.

Activation of the Ire1 RNase by oligomerization

We prepared variants of the cytosolic portion of Ire1 that contain the kinase and the RNase domains (Ire1KR), as well as the kinase and the RNase domains extended by 24 (Ire1KR24), 32 (Ire1KR32) or 120 (Ire1KR120, ref. 18) amino acids towards the amino terminus. These extensions are part of a ~120-amino-acid-long linker domain that tethers the kinase/RNase domains to the transmembrane region (Fig. 1a, c and Supplementary Fig. 1a). Ire1KR120 showed an RNase activity indistinguishable from that of Ire1KR32 but proved unsuitable for crystallization and was not pursued further. All Ire1 constructs site-specifically cleaved 5′-^{32}P-labelled stem-loop oligoribonucleotide[17] (HP21) derived from the *Xbp1* mRNA (Fig. 1b and Supplementary Fig. 1b; Methods). The observed rate constant exhibited a non-Michaelis dependence on the enzyme concentration and increased cooperatively with a Hill coefficient $n = 2$ for Ire1KR and Ire1KR24 and, surprisingly, a Hill coefficient $n = 3.5$–8 for Ire1KR32 (Fig. 1d). This observation indicates that the RNase activity of Ire1 arises from self-association with the formation of predominantly dimers for Ire1KR and Ire1KR24, and oligomers for Ire1KR32.

At protein concentrations above 10 μM, reactions with Ire1KR32 appeared as a heterogeneous suspension, indicating self-association of Ire1KR32 (Fig. 2a). The presence of several oligomeric species was apparent on analytical ultracentrifugation of the sample (Fig. 2b). The oligomerization could be readily reversed and RNase activity suppressed by addition of salt to the solution (Fig. 2a, c). The visible aggregation seemed to be specific because it was strongly induced by cofactors. In contrast, solutions of Ire1KR and Ire1KR24 remained clear at all concentrations with no signs of protein oligomerization, consistent with the lower cooperativity of their activation profiles (Fig. 1d).

[1]Department of Biochemistry and Biophysics, [2]Department of Cellular and Molecular Pharmacology, and [3]Howard Hughes Medical Institute, University of California at San Francisco, San Francisco, California 94158, USA. [4]Department of Molecular Cell and Developmental Biology at University of California, Santa Cruz, Santa Cruz, California 95064, USA.

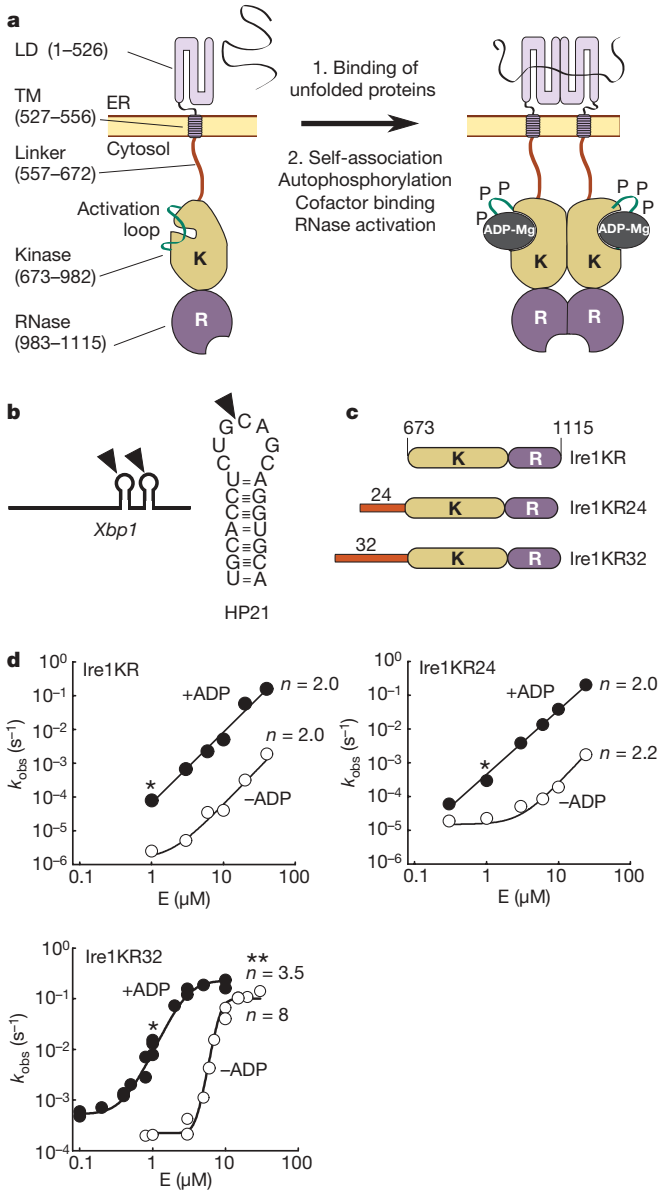


Figure 1 | Activation of Ire1 by self-association. **a**, A general scheme of Ire1 activation during the UPR summarizing the key events. The kinase domain of Ire1 is coloured light brown; the RNase domain is coloured purple. TM, transmembrane. **b**, Schematic representation of RNA substrates used in this work. Triangles mark sites of specific cleavage by Ire1. **c**, Ire1 constructs used for cleavage assays and structure determination. **d**, Cooperative activation profiles for Ire1KR, Ire1KR24 and Ire1KR32 obtained using 5′-^{32}P-HP21, with (filled circles) and without (open circles) cofactor. E, enzyme. Asterisks are used for reference in Fig. 2. Assay details are provided in Methods.

Ire1KR32 cleaved HP21 ~100-fold faster than did Ire1KR and Ire1KR24 (Fig. 2d, left panel). The observed rate constants were compared at 1 μM concentration of the enzymes, at which reactions occur in the same kinetic regime characterized by a log-linear concentration response of the observed rate constant. The catalytic advantage and the highly cooperative activation of Ire1KR32 were even more apparent with the *Xbp1* 443-base polymer, which more closely mimics the natural mRNA substrate of Ire1 (Fig. 2d, right panel, and Supplementary Fig. 2). These observations demonstrate that the N-terminal linker domain, particularly the eight amino acids that constitute the difference between Ire1KR32 and Ire1KR24 (Supplementary Fig. 1a), defines the self-association properties and the RNase activity of the cytosolic domains of Ire1. Notably, point mutations within these eight amino acids abrogate Ire1 signalling *in vivo*[19]. This functionally important linker extension was absent in the

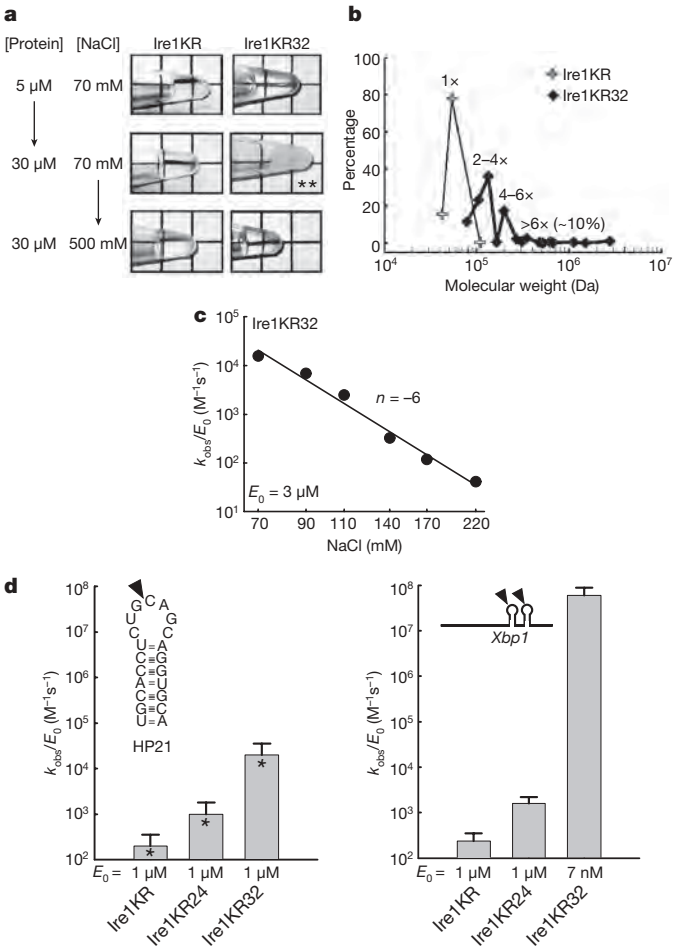


Figure 2 | Linker controls the oligomerization and activation of Ire1. **a**, Observation of visible self-association of Ire1KR32 that can be reversed by salt (NaCl). **b**, Analytical ultracentrifugation reveals monomers and dimers for Ire1KR as well as dimers and higher-order assemblies for Ire1KR32. Conditions were as in Fig. 1d; open symbols, 13.5 μM Ire1 (20 °C). **c**, Salt inhibits the RNase activity of Ire1KR32. **d**, Ire1KR32 has higher RNase activity against HP21 and *Xbp1* RNA compared to Ire1KR and Ire1KR24 (Supplementary Fig. 1b; 2). Error bars show variability between single-exponential fits from two to five independent measurements. Conditions similar to those used in Fig. 1d are marked * and **.

Ire1 variant used previously to obtain the back-to-back dimer structure[14]. The use of Ire1 that includes the extended N terminus resulted in the crystal structure of the oligomeric state of Ire1.

Structure of the Ire1 oligomer

To establish the mechanism of Ire1 oligomerization and activation, we used X-ray crystallography. Because efforts to crystallize Ire1KR32 with ADP produced crystals unsuitable for X-ray data collection, we attempted to co-crystallize Ire1KR32 with structurally diverse protein kinase inhibitors. Remarkably, several kinase inhibitors activated the RNase function revealing synthetic activators of wild-type Ire1 (Fig. 3a, b and Supplementary Fig. 3). These results have profound implications for therapeutic uses of kinase inhibitors (see Conclusions).

Crystals obtained with the inhibitor APY29 allowed determination of the structure of the Ire1KR32•APY29 complex at 3.9 Å resolution. The resolution improved to 3.2 Å with a mutant version of Ire1KR32, Ire1KR32Δ28•APY29, in which we deleted the αF–αEF loop (28 amino acids, 865–892). The αF–αEF loop is not evolutionary conserved and was disordered in the 3.9-Å structure. Its deletion had no effect on the RNase activity of Ire1 (Supplementary Fig. 4). Electron density for the APY29 molecule was found in the ATP-binding pocket of the Ire1 kinase domain (Fig. 3c). The position of APY29 indicates that it could form three hydrogen bonds with residues

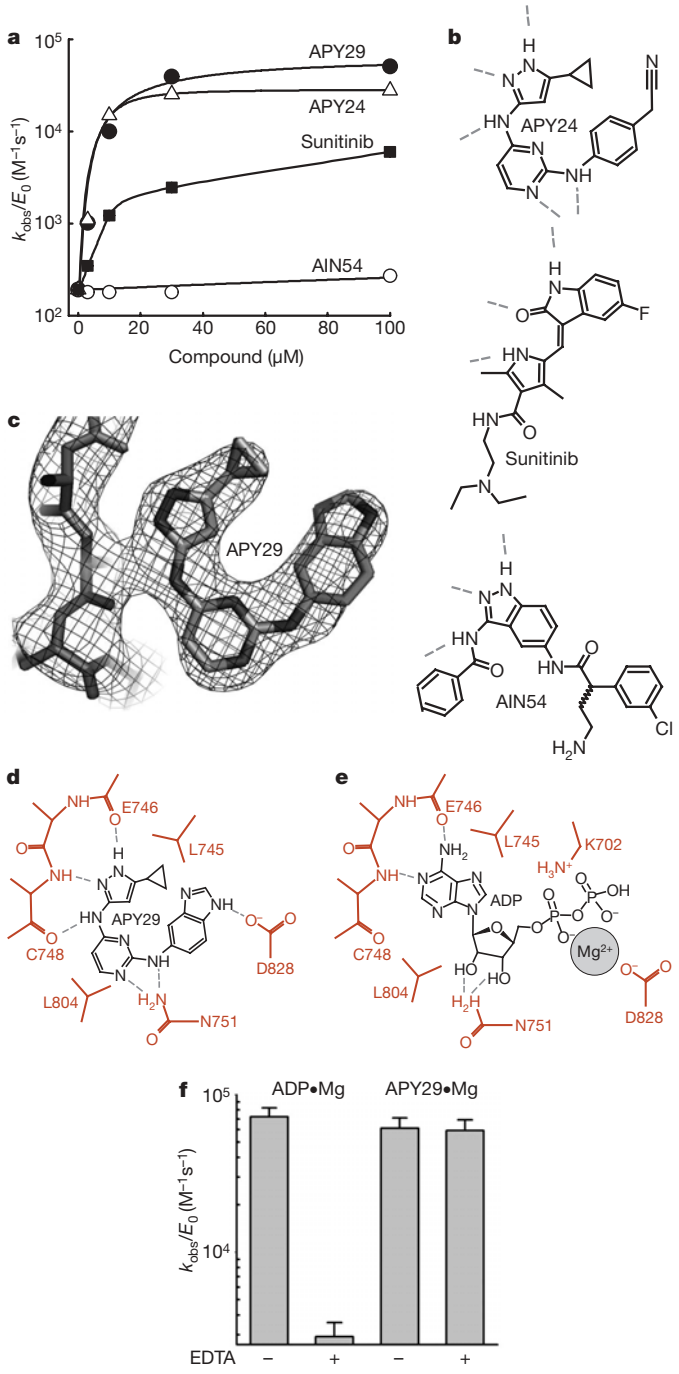


Figure 3 | Kinase inhibitors activate the RNase of wild-type Ire1. **a**, Activation of Ire1KR32 (3 μM) in the presence of different kinase inhibitors. **b**, Inhibitor structures, with probable hydrogen bonds shown by dashed lines. **c**, σ_A-weighted $3F_{obs}–2F_{calc}$ map for APY29 bound to Ire1KR32Δ28 contoured at 1.5σ. **d**, The network of probable hydrogen bonds between APY29 and Ire1. **e**, The network of interactions between ADP•Mg and Ire1 (PDB ID 2RIO). **f**, Chelation of magnesium inhibits Ire1 RNase in the presence of ADP, but not of APY29. Error bars show variability between single-exponential fits from two independent measurements. Reactions contained 2 mM ADP or 100 μM APY29.

Glu 746 and Cys 748 of the main chain and two additional hydrogen bonds or van der Waals contacts with the side chains Asn 751 and Asp 828 at the active site (Fig. 3d).

Although all tested compounds can potentially form hydrogen bonds with the protein backbone (Fig. 3b), the most potent activators, APY29 and APY24, also interact with the side chain Asn 751 and insert bulky aromatic rings in place of the ribose-phosphate moiety of ADP. Manual fitting of the FDA-approved anti-cancer drug Sunitinib guided by known structures of kinase•inhibitor complexes (Protein Data Bank (PDB) IDs 2G9X and 2F4J) predicts that the compound fills the adenine-binding site, but not the ribose and the phosphate subsites. Such partial occupancy could explain the fairly good binding of Sunitinib to Ire1 accompanied by partial activation of the enzyme (Fig. 3a). AIN54 could not be fit to the ATP pocket owing to steric clashes with the β1 strand.

The interactions of APY29 with the nucleotide-binding pocket closely mimic those of ADP except that APY29 does not use a divalent metal ion for docking (Fig. 3d, e). Accordingly, addition of EDTA inhibits the RNase activity of Ire1 for reactions stimulated by ADP but not APY29 (Fig. 3f). These findings support a model first proposed based on Ire1 mutants[16]: that ADP and kinase inhibitors activate Ire1 RNase by filling the ATP pocket. For maximum activity the adenine and the ribose subsites should be occupied, apparently to stabilize the active open conformation of the kinase that favours self-association of Ire1. Electrostatic interactions due to coordination of the magnesium ion and the phosphate groups of ADP do not have an indispensable role as the charged moieties can be replaced with neutral space-filling groups.

In contrast to Ire1, which lacks the oligomerization-inducing N-terminal segment and crystallizes as a back-to-back dimer[14], Ire1KR32 and Ire1KR32Δ28 crystallize as a symmetric high-order assembly (Fig. 4a, b). Fourteen Ire1 molecules constitute the asymmetric unit in the crystal lattice. Formation of the oligomer can be described by incremental addition of symmetric back-to-back Ire1 dimers to an end of a growing filament, with a simultaneous clockwise turn of 51.4° per dimer, with a complete 360° turn every 14 molecules.

The use of 14-fold non-crystallographic symmetry (NCS) improved the quality of averaged electron density maps and helped the modelling of all of the regions missing from the starting model (Supplementary Fig. 5). The structure of the kinase/RNase domain in the oligomer is similar to that in the Ire1•ADP dimer[14]. However, tight packing of Ire1 in the oligomer compared to the crystal packing of the Ire1•ADP dimers (Fig. 4a, inset) orders several fragments of Ire1 absent in the previous model (coloured green in Figs 4c–e). None of the new elements belong to the interface IF1^c defined previously in the back-to-back dimer[14] (Fig. 4b, c). Two new interfaces, IF2^c and IF3^c, form in addition to the interface IF1^c in the oligomer. Interface IF2^c has a two-fold symmetry and forms by contacts between the RNase domains of monomers A–D, C–F, and so on (Fig. 4d). Interface IF3^c creates a linear side-to-side arrangement of monomers into filaments (B → D → F and, with opposite polarity, A ← C ← E). Interface IF3^c is formed by contacts between the kinase domains and involves two new elements, the αD′ helix and the activation loop (Fig. 4e). The oligomerization-inducing N-terminal extension (residues 641–662) was disordered. Its structure and the mechanism of facilitating Ire1 oligomerization remain to be determined. It is possible that part of the N-tail contacts a dimerization interface, as proposed recently for the arginine-rich linker extension of epidermal growth factor receptor[20].

Architecturally, the oligomer resembles the double helix of DNA (Fig. 4b and Supplementary Fig. 6), where interface IF1^c parallels the interaction between nucleobases of opposing strands and interface IF3^c parallels phosphodiester linkages between nucleotides of the same strand.

The *trans*-autophosphorylation complex of Ire1

In the Ire1 oligomer, each kinase offers its activation loop to a new partner, thereby extending the filamentous oligomeric assembly. This interaction resembles the side-to-side arrangement of kinase dimers implicated in *trans*-autophosphorylation reactions[21,22]. Ire1KR32 used in this work contained 17 phosphorylated residues (Supplementary Fig. 7). Phosphorylation was not observed on expression of Ire1 with a kinase-inactivating mutation (D828A), indicating that all Ire1 phosphates derive from its own kinase activity

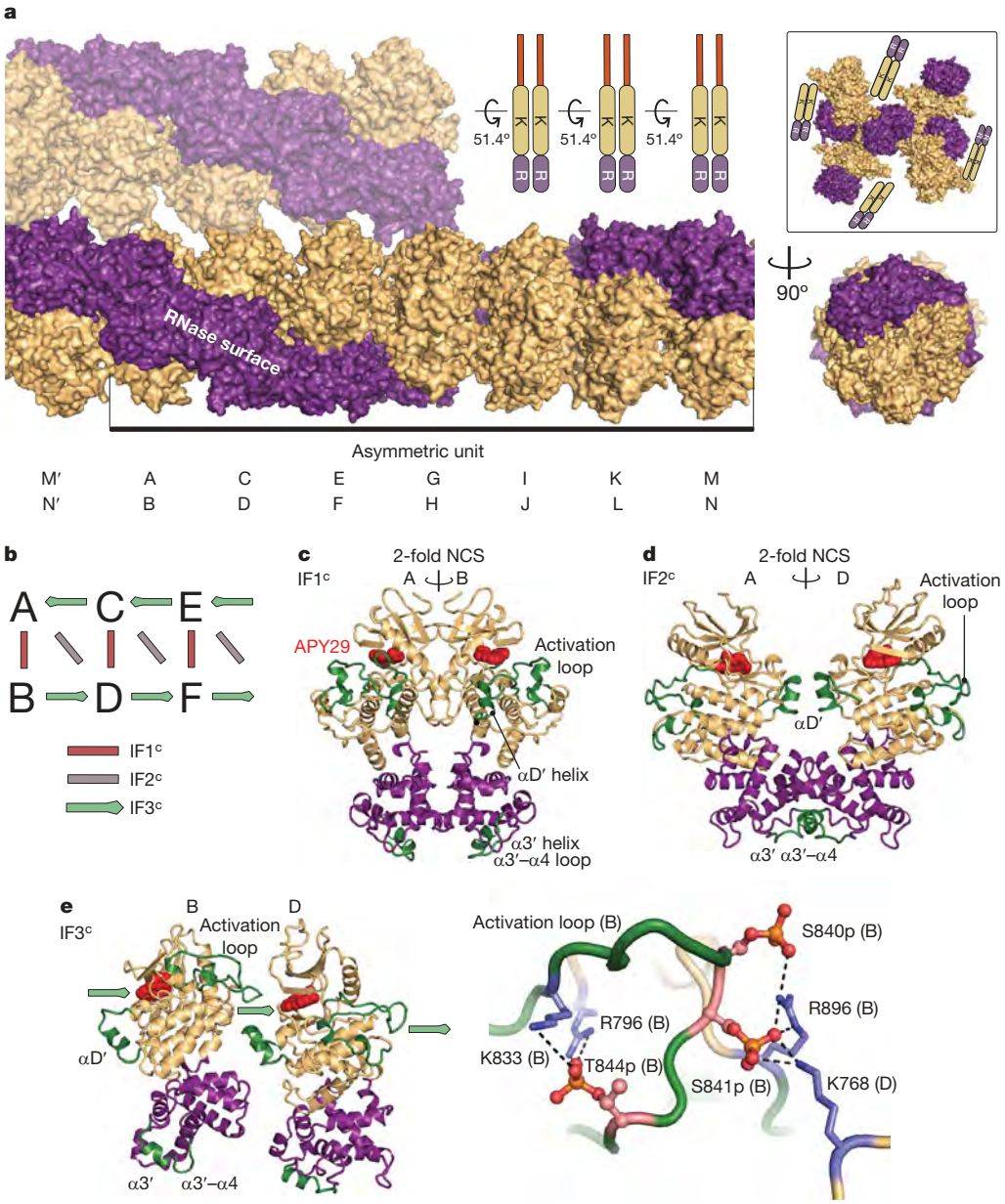


Figure 4 | Structure of the Ire1 oligomer. **a**, Assembly of Ire1KR32Δ28•APY29. A parallel filament in the crystal packing is shown above the main filament. Domains are coloured as in Fig. 1a. The inset shows crystal packing of Ire1 dimers (PDB ID 2RIO). Letters A–N below the structure refer to individual monomers in the asymmetric unit. K, kinase; R, ribonuclease. **b**, Three intermolecular interfaces of Ire1KR32Δ28 in the oligomer. **c**, Dimer formed via interface IF1^c. **d**, Dimer formed via interface IF2^c. **e**, Dimer formed via interface IF3^c (left). Close view of the activation loop (right). Phosphates are shown in ball representation. Arrows in **b** and **e** show the direction of the activation loop donation. Regions with previously unknown structures are coloured green.

as the protein is expressed in *Escherichia coli*. Mass spectrometric analyses localized the phosphorylation sites to the activation loop and the αEF–αF loop (Supplementary Table 1), both of which face interface IF3^c. Together, the oligomer structure and mass spectrometry support a model wherein IF3^c serves for transfer of the phosphates *in trans*, resolving the difficulty in explaining *trans*-autophosphorylation of Ire1 in the back-to-back dimer[14]. The tightly packed oligomer makes it highly unlikely that kinases other than Ire1 have access to the phospho-acceptor sites. This feature can explain the specific phosphorylation of sites in Ire1 that are not part of any recognizable consensus motif and the apparent absence of other kinases known to phosphorylate Ire1.

Three phosphorylated residues important for Ire1 activation *in vivo*[11]—Ser 840p, Ser 841p and Thr 844p—are resolved in the crystal structure (Fig. 4e). Thr 844p forms two intramolecular salt bridges positioned to stabilize the open state of the activation loop and conserved among kinases[23]. Ser 840p and Ser 841p form two additional intramolecular salt bridges, and Ser 841p forms a unique intermolecular salt bridge with an adjacent Ire1 molecule at interface IF3^c. All three phosphates are ideally placed to help Ire1 oligomerization by stabilizing the oligomerization-compatible open state of Ire1 kinase and positioning Ser 841p to stabilize interface IF3^c.

Three Ire1 interfaces control the RNase activity

The presence of the three distinct interfaces in the oligomer structure raised questions about their relative contribution to activation of the Ire1 RNase. Thus, we characterized Ire1 variants with each interface selectively impaired by mutations. For IF1^c, we prepared Ire1 with an E988Q mutation, which had the strongest deleterious effect on the RNase activity among the tested RNase IF1^c mutants[14]. For IF2^c and IF3^c, we identified previously uncharacterized contacts and mutated relevant residues to destabilize these contacts (Fig. 5a, b).

In the standard cleavage assay (3 μM Ire1, HP21 substrate), mutations at all three interfaces had significant deleterious effects (Fig. 5c). Mutations mapping to IF2^c and IF3^c exhibited as strong or stronger effects on the RNase activity as did the mutation mapping to IF1^c. Sedimentation profiles and activation profiles show that mutations at each of the interfaces weaken the self-association properties of Ire1KR32 (Supplementary Fig. 8). The functional importance of residues at all three interfaces indicates a conjoint effort from IF1^c, IF2^c and IF3^c in activation of the RNase.

The activation mechanism of the Ire1 RNase

It has been suggested that dimerization activates the RNase of Ire1 by complementing the pair of splice sites in the *HAC1*/*Xbp1* mRNA[14]. In principle, this mechanism could explain why the stem-loop HP21 is cleaved slowly compared to *HAC1*/*Xbp1* mRNA (Fig. 2d). However, several pieces of evidence do not support the proposed model. Only Ire1KR32 shows a large preference for cleavage of *HAC1*/*Xbp1* mRNA over stem-loop HP21, whereas Ire1KR and Ire1KR24 show no discrimination (Fig. 2d). Furthermore, we prepared a 354-nucleotide RNA substrate that contains only a single stem-loop but reacts with Ire1KR32 at the rate of *HAC1* and *Xbp1* mRNA (Fig. 6a, b). We also prepared a 58-nucleotide RNA substrate that contains two stem-loops but reacts with Ire1KR32 at a rate of HP21 (Fig. 6c). These findings show that substrates with dual stem-loops do not have catalytic advantage compared to substrates with a single stem-loop and that self-association activates the Ire1 RNase by a mechanism different from steric complementation of the dual splice sites. In the oligomer

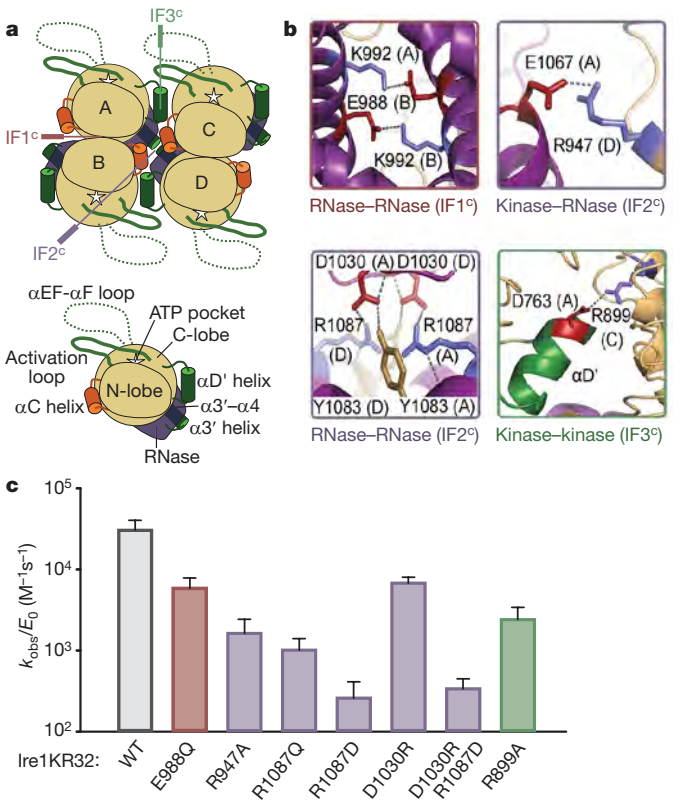


Figure 5 | Three interfaces of Ire1 contribute to the RNase activity. **a**, Schematic representation of the Ire1 oligomer packing. Stars mark the ATP-binding pocket of the Ire1 kinase. Lower panel shows a single monomer; upper panel shows packing of four monomers. **b**, Contacts at the intermolecular interfaces of the oligomer. **c**, Mutations of the predicted interface residues designed to weaken IF1^c, IF2^c and IF3^c inhibit the RNase activity of Ire1KR32. Reactions contained 5′-^{32}P-HP21, 3 μM Ire1KR32 and 2 mM ADP. WT, wild type. Error bars show variability between single-exponential fits from two independent measurements. The colour of the bars matches that in **a** and **b**.

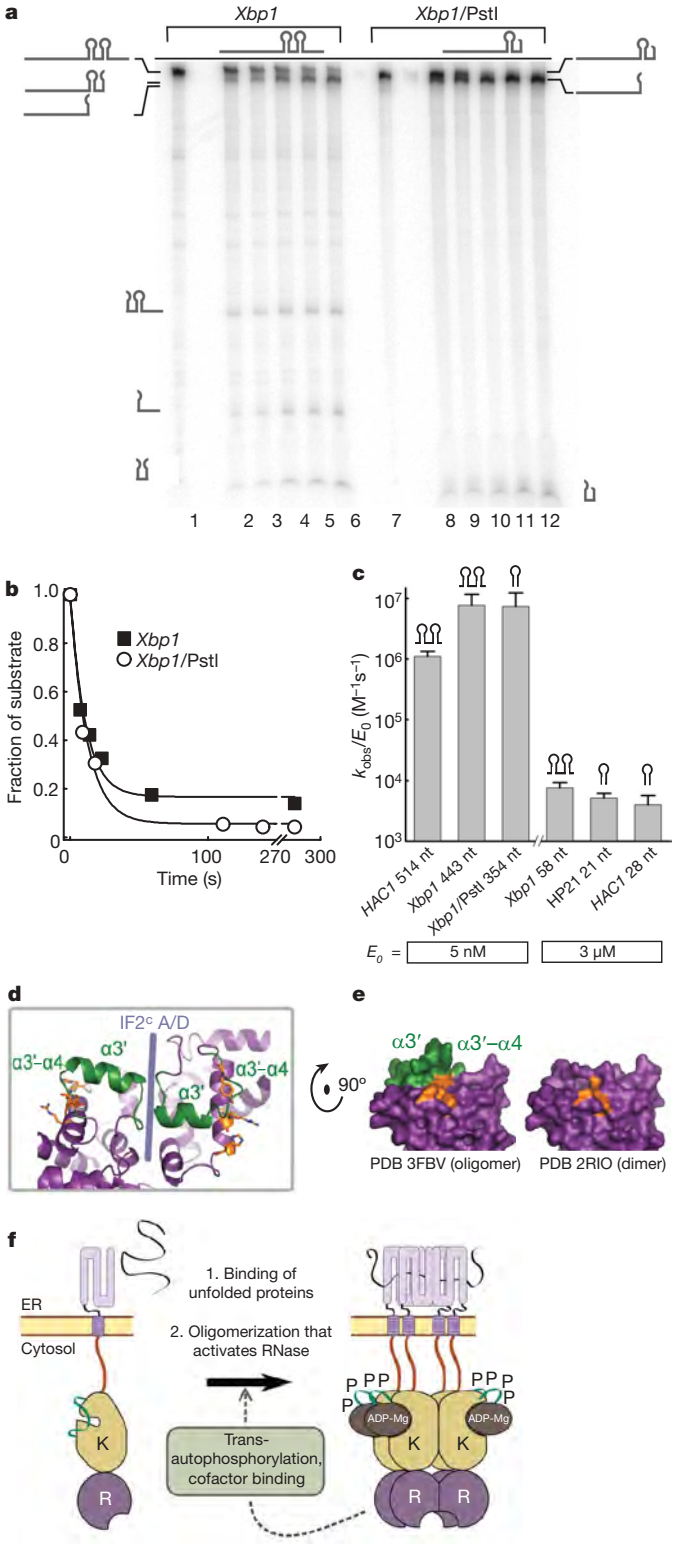


Figure 6 | The mechanism of Ire1 activation. **a**, Time courses for cleavage of *Xbp1* (lanes 1–6) and of *Xbp1*/PstI (lanes 7–12) with 5 nM Ire1KR32. **b**, Quantification of the gel in **a**. **c**, Cleavage of RNA with one and two splice sites by Ire1KR32. Error bars show variability between single-exponential fits from two to five independent measurements. nt, nucleotide. **d**, Position of the HLE at the RNase–RNase interface IF2^c. **e**, Molecular surface representation of the RNase domain in the oligomer and in the dimer structures. Putative catalytic residues in **d** and **e** are coloured orange; HLE is coloured green. **f**, Revised model of Ire1 activation during the UPR.

structure, the RNase domains are linked into a continuous ribbon by two interfaces, IF1^c and IF2^c (Supplementary Fig. 9). IF2^c places the α3′ helices from the adjacent RNase monomers in reciprocal contact (Fig. 6d). The α3′ helix as well as the adjacent α3′–α4 loop (coloured green in Fig. 6d, e and Supplementary Fig. 9) are disordered in the back-to-back dimer structure[14], indicating that IF2^c stabilizes this helix-loop element (designated HLE). Location of HLE near a dimerization interface may provide a dynamic switch that controls the RNase activity of Ire1. Indeed, HLE completes the proposed RNase active site and creates a cavity characteristic for substrate-binding pockets of enzymes (Fig. 6e), and point mutations in the HLE inactivate the RNase[14] .

An important additional contribution of oligomerization to the RNase activation may result from the extensive molecular surface of the oligomer, which would provide interactions with the substrate mRNA not possible with a monomer or a dimer. *Xbp1* mRNA binds to Ire1KR32 several orders of magnitude better than HP21 (Supplementary Fig. 10), indicating that protein–mRNA binding interactions spread beyond a stem-loop binding site of Ire1. These interactions could explain why Ire1KR and Ire1KR24, which do not form oligomers, do not discriminate between large and small RNA substrates (Fig. 2d). We conclude that yet to be characterized extended contacts between the Ire1 oligomer and mRNA feature prominently in Ire1 function.

Mechanistic implications

Key attributes of Ire1 activation emerged soon after its discovery and include Ire1 self-association, *trans*-autophosphorylation and the binding of ADP as a cofactor (Fig. 1a). Our present functional and structural data rationalize each of these events. In particular, we show that the primary step activating the Ire1 RNase is the self-assembly of the cytosolic region into a helical rod structure (Fig. 4a). The self-association equilibrium built into the cytosolic kinase/RNase module must be subservient to the ligand-controlled oligomerization of the ER-lumenal domain of Ire1 to establish the flow of the UPR signal from the ER lumen towards the cytosol. Therefore, aggregation of the ER-lumenal domain of Ire1 by unfolded proteins would serve to increase the local concentrations of the kinase/RNase domains on the cytosolic side of the ER membrane, passing the threshold for oligomerization and, consequently, RNase activation.

The roles and the temporal separation of *trans*-autophosphorylation and ADP binding are now clear. Oligomerization of the unphosphorylated Ire1 opens the kinase domain and positions it for *trans*-autophosphorylation (Supplementary Fig. 11a). ATP enters the opened kinase and phosphorylates the activation loop *in trans* to lock it in the oligomerization-compatible open state and to introduce a phosphate-mediated salt bridge at the interface IF3^c. These events provide positive feedback for oligomer assembly (Fig. 6f). Binding of a cofactor occurs in the open state of Ire1 kinase, shifts the equilibrium from monomers towards multimers and provides an additional, phosphorylation-independent level of positive modulation for the activating transition (Fig. 6f). At increased concentrations, Ire1 self-associates and becomes activated independent of phosphorylation (Supplementary Fig. 11b) and cofactor binding (Fig. 1d, open circles), directly supporting the model wherein oligomerization is the earliest and centremost step of Ire1 activation. Cofactor binding and phosphorylation enhance the self-association properties of Ire1 but neither is strictly required.

A model for the structure of the UPR-induced Ire1 foci[12,13] emerges from our work. Oligomers formed by the ER-lumenal domain of Ire1 and the cytoplasmic domains can be arranged to give similar periodicity of monomers on both sides of the ER membrane (Supplementary Fig. 12a). The resulting mesh could provide a platform for the formation and growth of supramolecular Ire1 foci in two dimensions. The length of the linkers connecting the functional domains of Ire1 to the transmembrane region permits this arrangement (Supplementary Fig. 12b). Such an assembly would allow a cooperative response to unfolded proteins and a prolonged time to mount and extinguish the UPR.

Conclusions

Our knowledge of the multi-domain signalling proteins built around protein kinases is fairly immature. This work shows how one of these molecules, Ire1, operates by forming a supramolecular structure not observed previously. We found that kinase inhibitors—including Sunitinib—act as potent activators of the Ire1 RNase. Because Ire1 provides cytoprotective function[24,25] from which cancer cells may benefit, it may be of therapeutic value to separate the intended function of kinase inhibitors towards the targets for which they were designed from activation of Ire1. Conversely, Ire1 activation might contribute to the beneficial effects of kinase inhibitors, including Sunitinib, in mouse models of type 1 diabetes[26], and the cytoprotective effect of the Ire1 activators may be harnessed to combat protein-folding diseases.

METHODS SUMMARY

Proteins were expressed in *E. coli* and purified using glutathione *S*-transferase (GST)-affinity purification and size-exclusion chromatography. The DNA oligonucleotides were made by PCR or purchased from IDT. RNA oligonucleotides were purchased from Dharmacon Inc. or prepared by *in vitro* transcription with T7 RNA polymerase. All kinetic assays were done at 30 °C and neutral pH. Diffraction data were collected from cryo-preserved crystals at a beamline 8.3.1 (Advanced Light Source, Berkeley National Laboratories). The structure was solved at 3.2 Å resolution by molecular replacement followed by refinement in CNS[27] and PHENIX[28]. The 2.4-Å structure of Ire1 dimer (PDB ID 2RIO[14]) was used as a molecular replacement search model in PHASER[29]. The final model containing amino acids 663–864 and 893–1115 of Ire1 has R/R_{free} of 0.235/0.283 and excellent stereochemistry (PDB ID 3FBV; Supplementary Fig. 13 and Supplementary Tables 2–5). A part of the N-tail (residues 641–662) is disordered.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 May; accepted 25 November 2008.
Published online 14 December 2008.

1. Cox, J. S. & Walter, P. A novel mechanism for regulating activity of a transcription factor that controls the unfolded protein response. *Cell* **87,** 391–404 (1996).
2. Yoshida, H., Matsui, T., Yamamoto, A., Okada, T. & Mori, K. XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell* **107,** 881–891 (2001).
3. Koong, A. C., Chauhan, V. & Romero-Ramirez, L. Targeting XBP-1 as a novel anti-cancer strategy. *Cancer Biol. Ther.* **5,** 756–759 (2006).
4. Ma, Y. & Hendershot, L. M. The role of the unfolded protein response in tumour development: friend or foe? *Nature Rev. Cancer* **4,** 966–977 (2004).
5. Zheng, Y. *et al.* Hepatitis C virus non-structural protein NS4B can modulate an unfolded protein response. *J. Microbiol.* **43,** 529–536 (2005).
6. Kudo, T. *et al.* The unfolded protein response is involved in the pathology of Alzheimer's disease. *Ann. NY Acad. Sci.* **977,** 349–355 (2002).
7. Bartoszewski, R. *et al.* Activation of the unfolded protein response by {Delta}F508 CFTR. *Am. J. Respir .Cell. Mol. Biol.* **39,** 448–457 (2008).
8. Naidoo, N., Giang, W., Galante, R. J. & Pack, A. I. Sleep deprivation induces the unfolded protein response in mouse cerebral cortex. *J. Neurochem.* **92,** 1150–1157 (2005).
9. Atkin, J. D. *et al.* Endoplasmic reticulum stress and induction of the unfolded protein response in human sporadic amyotrophic lateral sclerosis. *Neurobiol. Dis.* **30,** 400–407 (2008).
10. Credle, J. J., Finer-Moore, J. S., Papa, F. R., Stroud, R. M. & Walter, P. On the mechanism of sensing unfolded protein in the endoplasmic reticulum. *Proc. Natl Acad. Sci. USA* **102,** 18773–18784 (2005).
11. Shamu, C. E. & Walter, P. Oligomerization and phosphorylation of the Ire1p kinase during intracellular signaling from the endoplasmic reticulum to the nucleus. *EMBO J.* **15,** 3028–3039 (1996).
12. Kimata, Y. *et al.* Two regulatory steps of ER-stress sensor Ire1 involving its cluster formation and interaction with unfolded proteins. *J. Cell Biol.* **179,** 75–86 (2007).
13. Aragón, T. *et al.* Messenger RNA targeting to endoplasmic reticulum stress signalling sites. *Nature* doi:10.1038/nature07641 (this issue).
14. Lee, K. P. *et al.* Structure of the dual enzyme Ire1 reveals the basis for catalysis and regulation in nonconventional RNA splicing. *Cell* **132,** 89–100 (2008).
15. Zhang, X., Gureasko, J., Shen, K., Cole, P. A. & Kuriyan, J. An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* **125,** 1137–1149 (2006).
16. Papa, F. R., Zhang, C., Shokat, K. & Walter, P. Bypassing a kinase activity with an ATP-competitive drug. *Science* **302,** 1533–1537 (2003).

17. Gonzalez, T. N. & Walter, P. Ire1p: a kinase and site-specific endoribonuclease. *Methods Mol. Biol.* **160,** 25–36 (2001).
18. Sidrauski, C. & Walter, P. The transmembrane kinase Ire1p is a site-specific endonuclease that initiates mRNA splicing in the unfolded protein response. *Cell* **90,** 1031–1039 (1997).
19. Goffin, L. *et al.* The unfolded protein response transducer Ire1p contains a nuclear localization sequence recognized by multiple beta importins. *Mol. Biol. Cell* **17,** 5309–5323 (2006).
20. Thiel, K. W. & Carpenter, G. Epidermal growth factor receptor juxtamembrane region regulates allosteric tyrosine kinase activation. *Proc. Natl Acad. Sci. USA* **104,** 19238–19243 (2007).
21. Pirruccello, M. *et al.* A dimeric kinase assembly underlying autophosphorylation in the p21 activated kinases. *J. Mol. Biol.* **361,** 312–326 (2006).
22. Pike, A. C. *et al.* Activation segment dimerization: a mechanism for kinase autophosphorylation of non-consensus sites. *EMBO J.* **27,** 704–714 (2008).
23. Yonemoto, W. *et al.* Autophosphorylation of the catalytic subunit of cAMP-dependent protein kinase in *Escherichia coli*. Identification of phosphorylation sites in the recombinant catalytic subunit of cAMP-dependent protein kinase. *Protein Eng.* **10,** 915–925 (1997).
24. Lin, J. H. *et al.* IRE1 signaling affects cell fate during the unfolded protein response. *Science* **318,** 944–949 (2007).
25. Han, D, *et al.* A kinase inhibitor activates the IRE1α RNase to confer cytoprotection against ER stress. *Biochem. Biophys. Res. Commun.* **365,** 777–783 (2008).
26. Louvet, C. *et al.* Tyrosine kinase inhibitors reverse type 1 diabetes in nonobese diabetic mice. *Proc. Natl Acad. Sci. USA* **105,** 18895–18900 (2008).
27. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
28. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).
29. McCoy, A. J. Phaser crystallographic software. *J. Appl. Cryst.* **40,** 658–674 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Krutchinsky for the help with MALDI instruments and for the tryptic digest analysis of Ire1KR32, F. Gruswitz for useful discussions, C. Waddling for managing the protein crystallization facility of the molecular structure group (MSG) at UCSF, and to the staff of the beamline 8.3.1 at the Advanced Light Source (Berkeley). We thank members of the Walter laboratory for critical review of the manuscript. A.V.K. is a recipient of Jane Coffin Childs fellowship. C.Z. was supported by a grant from the National Parkinson Foundation. R.M.S., J.F.-M. and P.F.E. were supported by an NIH grant RO1 GM60641. P.W. and K.M.S. are Investigators of the Howard Hughes Medical Institute.

Author Contributions A.V.K. designed and prepared protein and RNA constructs and carried out kinetic and biophysical analyses. A.V.K. and P.F.E. carried out crystallization and data collection. A.A.K. performed structure determination. J.F.-M. and A.V.K. contributed to crystallographic data processing and model building. C.Z. and K.M.S. selected and provided the kinase inhibitors. P.W. and R.M.S. supervised the work. A.V.K. and P.W. wrote the manuscript.

Author Information Atomic coordinates and structure factors for the reported crystal structure have been deposited in the Protein Data Bank under accession number 3FBV. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.V.K. (alexei.korennykh@ucsf.edu).

METHODS

Experimental errors. All quantitative parameters were measured two or more times. The rate variations between measurements done on different days were within twofold and were always small compared to the effects we describe as significant. Errors and experimental uncertainties are indicated where applicable.

Expression and purification of Ire1 constructs. The plasmids for Ire1 expression were prepared using PCR with Pfu polymerase and pGEX-6P-2 vector encoding the cytoplasmic domain of Ire1 (ref. 18). DNA primers for mutagenesis were designed using Biochem Lab Solutions 3.5 and purchased from IDT. Proteins were expressed in BL21 CodonPlus (RIPL) *E. coli* cells (Stratagene). Expression and purification was conducted as described previously[30]. Bacteria were grown at 22 °C and lysis and FPLC buffers contained at least 300 mM NaCl to prevent aggregation of Ire1. Protein concentrations were determined from ultraviolet spectra using absorption peak at 280 nm and calculated extinction coefficients (Biochem Lab Solutions 3.5). Stocks of purified Ire1 had concentrations 10–70 mg ml^{-1} and were at least 99% pure as judged by Coomassie blue staining and quantification of FPLC traces.

Preparation of RNA substrates. HP21 21- and *HAC1* 28-base polymers were purchased from Dharmacon Inc. Other RNA substrates were prepared from restriction-digested plasmids encoding *HAC1* and *Xbp1* mRNA or from PCR-amplified products. Preparative amounts of long RNA were made using MegashortScript kit (Ambion). Before use, the oligonucleotides were purified by a denaturing (8 M urea) 5–20% polyacrylamide gel electrophoresis (PAGE). Cross-linked 29:1 polyacrylamide (40%) was purchased from National Diagnostics. Gel slices containing RNA were eluted in TE buffer and ethanol-precipitated. Substrates labelled with ^{32}P at the 5′-terminus were prepared using T4 PNK (NEB) and γ-^{32}P-ATP (Perkin Elmer). The ^{32}P-body-labelled substrates were prepared by transcription with T7 RNA polymerase (Promega) in the presence of α-^{32}P-UTP (Perkin Elmer). All ^{32}P-labelled substrates were purified by denaturing 5–20% PAGE, eluted in TE and ethanol-precipitated before using.

Preparation of kinase inhibitors. Kinase inhibitors were obtained by chemical synthesis. The synthetic schemes will be reported in an upcoming publication.

The RNase cleavage assay. RNA cleavage reactions were conducted at 30 °C in buffer containing 20 mM HEPES (pH 7.0 at 30 °C), 70 mM NaCl, 2 mM ADP (pH 7.0), 2 mM $Mg(OAc)_2$, 5 mM DTT, 5% glycerol, less than 1 nM ^{32}P-labelled RNA substrate and 3 nM to 20 μM Ire1. Ire1 cleaves *HAC1*- and *Xbp1*-derived RNA substrates with similar kinetics (Fig. 6b). *Xbp1* mRNA reacts to lower end points and has a more robust kinetic behaviour compared to that of *HAC1* mRNA, presumably owing to better folding, and was used in most of the experiments. Reaction solutions and buffers were designed using Biochem Lab Solutions 3.5. Reactions were prepared such that 1 μl of RNA was added to 9 μl of pre-warmed reaction mixture containing all components except RNA. Typically, 3–10-min time courses were collected starting from 5 s for the first time point. At time intervals, 1 μl aliquots were withdrawn from each reaction and mixed with 6 μl stop solution containing 10 M urea, 0.1% SDS, 0.1 mM EDTA, 0.05% xylene cyanol and 0.05% bromophenol blue. The samples were separated by a denaturing 10–15% PAGE and exposed on a phosphor storage screen. The screens were scanned on a Storm or a Typhoon instruments and quantified using ImageQuant 5.0 or GelQuant.NET 1.4 programs. The data were plotted and fit in SigmaPlot 6.1. Hill equation with one added constant that describes the low-enzyme plateau was used to fit the cooperative activation profiles.

Analytical ultracentrifugation. Ire1 samples (13.5 μM) were loaded to a 400 μl cell in buffer for the RNase cleavage assay. Centrifugation was carried out on a Beckmann XL-A analytical ultracentrifuge at 129,024*g* (20 °C). A total of 60–80 scans were collected. Sedimentation traces were analysed in UltraScan 9.8 using C(s) model.

Mass spectrometry. A gold-coated plate was washed with 100% methanol and water. Solution A (10 mg 4-HCCA in 0.7 ml of acetonitrile, 0.1% trifluoroacetic acid) was quickly spread in a layer and allowed to dry. The residue was removed gently with a tissue. 0.5 μl of a mixture containing 1 μl Ire1 sample (0.1–1 mg ml^{-1}) and 5 μl solution B (300 μl formic acid, 100 μl H_2O, 200 μl iso-propanol, 10 mg 4-HCCA) was spotted over the dried surface and allowed to dry. The sample was washed twice with 2 μl of 0.1% trifluoroacetic acid and used for MALDI analysis on a Voyager mass spectrometer. The spectra were analysed using MoverZ (Genomic Solutions).

Crystallization. Initially Ire1KR32 (10 mg ml^{-1}) was crystallized as a complex with ADP (2 mM) by vapour diffusion in hanging drops from 1.0 M sodium citrate. These co-crystals were disordered in one direction preventing their use in diffraction studies. A different crystal form was found by replacing the ADP with a kinase inhibitor APY29. Ire1KR32•APY29 crystals were also obtained by vapour diffusion in hanging drops. Drops were prepared by mixing 1 μl of Ire1KR32 (12 mg ml^{-1}) and APY29 (1 mM) in buffer containing 20 mM HEPES, pH 7.0 (20 °C), 500 mM NaCl, 2 mM DTT and 5% glycerol with 1 μl of solution containing 0.27 M Na_2SO_4, 8% PEG-3350, 10 mM EDTA and 2 mM TCEP. Well solution contained 200 μl of 0.085 M Na_2SO_4, 2.33% PEG-3350 and 5% tert-amyl alcohol. Single crystals grew at room temperature (20 °C) to a maximum size of $0.1 \times 0.4 \times 0.2$ mm^3 during three to four days. For data collection, the crystals were flash-frozen in solution containing 0.085 M Na_2SO_4, 3% tert-amyl alcohol, 5% PEG-3350 and 30% ethylene glycol. Crystals of Ire1KR32Δ28, were grown as described previously, except well solution contained 90 mM Na-citrate, pH 5.6.

Data collection and analysis. X-ray diffraction data were recorded on a beam line BL 8.3.1 at the Advanced Light Source (Berkeley National Laboratory). The data set, obtained using an X-ray wavelength of 1.11587 Å and an oscillation angle of 1 degree, was indexed and integrated using the XDS package[31] (Supplementary Table 2). The data set was scaled with SCALA[32]. Five per cent of the reflections were marked as a test set. To reduce possible bias in R_{free} towards R due to the high NCS, we have selected test-set reflections in thin shells rather than randomly[33], so that the symmetry-related reflections belong either to the test or to the working set. A molecular replacement solution was found using PHASER[29]. Fourteen copies of monomer A from the X-ray structure of the Ire1 dimer[14] were used as a starting model for refinement. Simulated annealing and grouped B-factor refinement were carried out in CNS[27] followed by simulated annealing (starting at 10,000K) and TLS refinement in PHENIX[28], using 14-fold NCS. One NCS group was comprised of one Ire1 monomer. HLE of monomer A was excluded from NCS treatment because its conformation differed from that in other monomers due to a close crystallographic contact. Fourier σ_A-weighted[34] F_{obs}–F_{calc} difference maps were used for interpretation of the parts of the model missing from the starting structure. The model of the ligand was created using ChemSketch 10.0 from Advanced Chemistry Development, Inc. (ACD/Labs). Model building and local real-space refinement were performed in Coot[35], PyMol (DeLano Scientific) and RSRef[36]. The resulting model has excellent stereochemical parameters (Supplementary Table 2), no Ramachandran plot outliers (0 residues in disallowed regions) and low crystallographic R/R_{free} of 0.235/0.283, indicating good agreement with diffraction data.

30. Nock, S., Gonzalez, T. N., Sidrauski, C., Niwa, M. & Walter, P. Purification and activity assays of the catalytic domains of the kinase/endoribonuclease Ire1p from *Saccharomyces cerevisiae*. *Methods Enzymol.* **342**, 3–10 (2001).
31. Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Cryst.* **26**, 795–800 (1993).
32. Evans, P. Scaling and assessment of data quality. *Acta Crystallogr. D* **62**, 72–82 (2006).
33. Fabiola, F., Korostelev, A. & Chapman, M. S. Bias in cross-validated free *R* factors: mitigation of the effects of non-crystallographic symmetry. *Acta Crystallogr. D* **62**, 227–238 (2006).
34. Read, R. J. Coefficients for maps using phases from partial structures with errors. *Acta Crystallogr. A* **42**, 140–149 (1986).
35. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
36. Korostelev, A., Bertram, R. & Chapman, M. S. Simulated-annealing real-space refinement as a tool in model building. *Acta Crystallogr. D* **58**, 761–767 (2002).

ARTICLES

Visualization of a missing link in retrovirus capsid assembly

Giovanni Cardone[1], John G. Purdy[2], Naiqian Cheng[1], Rebecca C. Craven[2] & Alasdair C. Steven[1]

For a retrovirus such as HIV to be infectious, a properly formed capsid is needed; however, unusually among viruses, retrovirus capsids are highly variable in structure. According to the fullerene conjecture, they are composed of hexamers and pentamers of capsid protein (CA), with the shape of a capsid varying according to how the twelve pentamers are distributed and its size depending on the number of hexamers. Hexamers have been studied in planar and tubular arrays, but the predicted pentamers have not been observed. Here we report cryo-electron microscopic analyses of two *in-vitro*-assembled capsids of Rous sarcoma virus. Both are icosahedrally symmetric: one is composed of 12 pentamers, and the other of 12 pentamers and 20 hexamers. Fitting of atomic models of the two CA domains into the reconstructions shows three distinct inter-subunit interactions. These observations substantiate the fullerene conjecture, show how pentamers are accommodated at vertices, support the inference that nucleation is a crucial morphologic determinant, and imply that electrostatic interactions govern the differential assembly of pentamers and hexamers.

A retrovirus has a lipoprotein envelope lined with a layer of matrix protein (MA), surrounding a nucleoprotein core[1]. In the core, the diploid RNA genome in complex with nucleocapsid protein (NC) and the replication enzymes is enclosed within the capsid—a shell of CA protein. MA, CA and NC are derived from a common precursor, the Gag polyprotein, which assembles into a thick-walled spherical shell in the immature virus. After it buds off from the host cell, the viral protease is activated, releasing CA subunits that assemble into capsids. Capsids of a given retrovirus vary in structure, and the predominant types vary among retroviruses[1]; for instance, those of HIV are conical[2], those of Rous sarcoma virus (RSV) are irregular polyhedra[3,4], and those of murine leukemia virus are roundish[5]. Some virions contain more than one capsid, and nested (that is, multilayer) capsids are also observed[2,4,6,7].

The CA subunit has an amino-terminal domain (NTD) and a carboxy-terminal domain (CTD), connected by a flexible linker. High-resolution structures have been determined for both domains for several retroviruses[8–16]. Although there is minimal sequence conservation except in the major homology region (MHR)—a 20-residue tract in the CTD—the two folds are conserved. Both are highly α-helical; NTD has seven helical segments (helices 1–7) and CTD has four (helices 8–11). So far, no retroviral capsid has been solved at high resolution, but progress has been made in electron microscopy studies of *in-vitro*-assembled sheets and tubes[17–21], which consist of NTD hexamers connected by dimerization of adjacent CTDs.

A plausible model for conical HIV capsids was derived by generalizing the 'fullerene' architecture that underlies the icosahedral capsids of many viruses, which have 12 evenly distributed pentamers interspersed by hexamers[22]. Differently sized capsids are distinguished by triangulation numbers (T) in the sequence {1, 3, 4, 7, …}, with a capsid having 12 pentamers and $10(T-1)$ hexamers[23]. Whereas an icosahedral capsid has both ends capped with six pentamers in symmetric (5 + 1) configuration, cones are envisaged to have five pentamers at their narrow ends and seven at their wide ends[17,22,24], and 'angular' RSV capsids to have six pentamers at both ends, distributed less regularly than in an icosahedron[4,25]. However, no pentamers have been visualized so far.

Assembly of CA into small isometric capsids

In vitro assembly of full-length CA protein of RSV produces a diversity of structures, including spheroids, tubes and planar arrays[14]. In buffer containing 0.5 M phosphate at near-neutral pH, the protein can also form angular structures, resembling capsids inside native virions[26]. However, a major portion of the assemblies formed in 0.5 M phosphate is small isometric particles (Fig. 1a). Most are ~17 nm in diameter; a few are larger, at ~30 nm. The 17-nm particles indicate an interpretation of the small rings with the wall thickness of capsids that were observed in tomographic slices of RSV virions (Fig. 3b of ref. 16); that is, the rings are likely to represent slices through $T = 1$ capsids. It seems, therefore, that such capsids, although too small to confine genomes, are also assembled *in situ*.

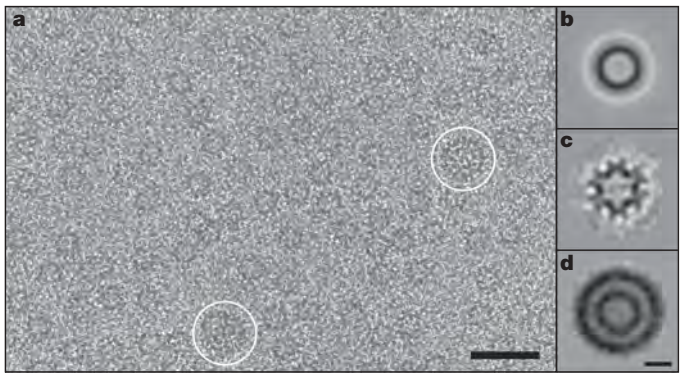


Figure 1 | RSV CA protein assembles *in vitro* in 0.5 M phosphate buffer into small isometric particles. a, Cryo-electron micrograph of capsids; most are ~17 nm in diameter, whereas a few are ~30 nm (white circles). Scale bar, 50 nm. **b–d**, Averaged images. **b**, 17-nm capsids, unclassified. **c**, 17-nm capsids, projecting the three-fold view. **d**, 30-nm capsids, unclassified, showing two concentric shells. Scale bar, 10 nm.

[1]Laboratory of Structural Biology, National Institute for Arthritis, Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. [2]Department of Microbiology and Immunology, The Pennsylvania State University College of Medicine, Hershey, Pennsylvania 17033, USA.

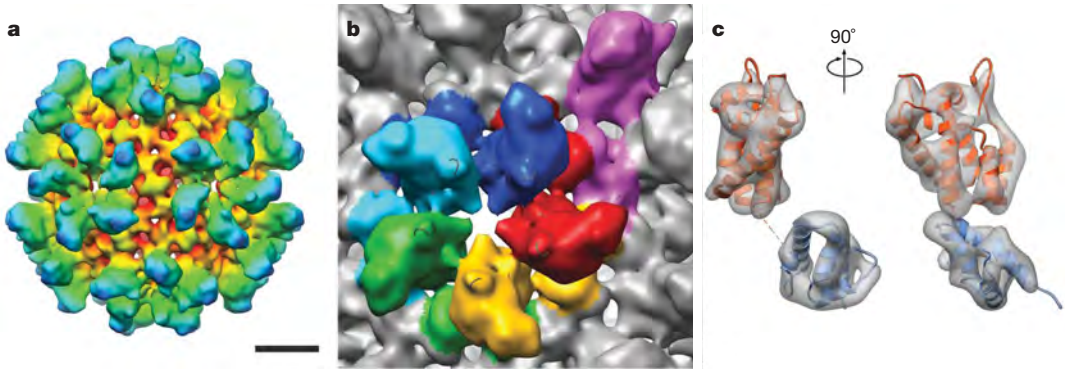


Figure 2 | Three-dimensional reconstruction of the RSV CA $T = 1$ capsid. **a**, Surface rendering coloured radially, red to blue, viewed along a two-fold axis of symmetry. Scale bar, 5 nm. **b**, Segmented view of a pentamer and surrounds, coloured by subunit. An additional subunit from an adjacent pentamer is purple. The centre-to-centre spacing between pentamers is 8.7 nm at mid-floor. **c**, Two views of a single subunit (left and right), with fitted pseudo-atomic model. The density is contoured at 3σ. Fitted were an NTD (orange, residues 1 to 147) and a CTD (blue, residues 152 to 230).

The similar polymorphisms displayed *in vitro* and *in situ* imply that form determination is largely an intrinsic property of CA protein, although other factors may steer assembly in certain directions—for instance, by affecting nucleation[4].

The capsids are icosahedrally symmetric, $T = 1$ and $T = 3$

The 17-nm particles have a high-density rim (Fig. 1a, b), indicating that they are hollow spheres. Classification of the images detected subsets with two-, three- and five-fold symmetry, consistent with them being projections of an icosahedron. Accordingly, we used the three-fold view (Fig. 1c), which was the most populated and thus had the lowest noise level, to generate a three-dimensional reconstruction. The result (not shown) strongly indicated that this is a $T = 1$ capsid. This reconstruction was then refined by projection-matching to <1.0-nm resolution (Fig. 2a, b). It shows a capsid consisting of 12 pentamers. Five protruding domains surround each vertex in crown-like structures, giving a full particle diameter of 23 nm. The mean diameter in the middle of the 'floor' layer of density is 16 nm. Each pentamer has a small axial hole, ~1.2 nm across.

Most 30-nm particles have an inner as well as an outer shell (Fig. 1d). Although their scarcity limited the scope for reconstruction, these data yielded a density map at 2.2-nm resolution. It shows the outer shell to be a $T = 3$ capsid with 12 pentamers and 20 hexamers, surrounding a $T = 1$ capsid (Fig. 3a, b). The two shells are in register, with coaxial stacking of their vertices and a relative rotation of ~36° between the respective pentamers. The full diameter is 35 nm and the mean diameter in the floor is 28 nm.

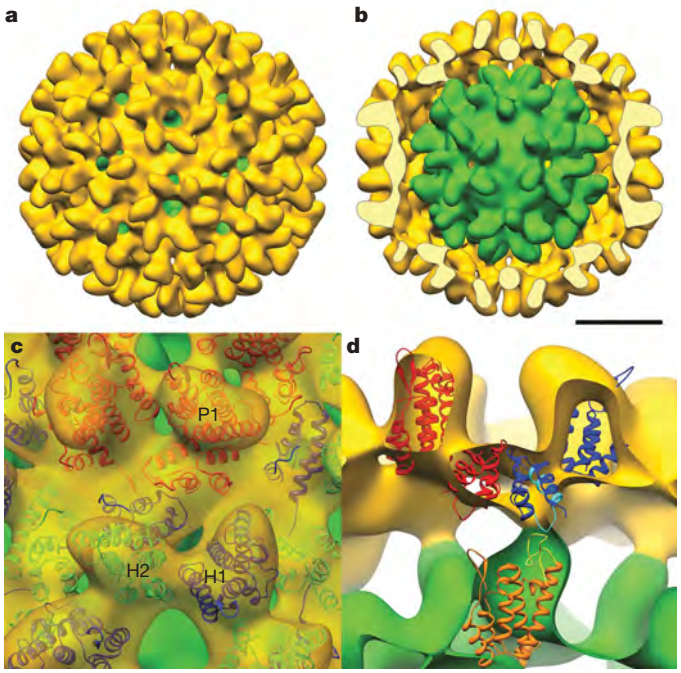


Figure 3 | The 30-nm RSV CA double-layer capsid. **a**, Surface rendering of cryo-electron microscopy reconstruction showing the outer $T = 3$ capsid (yellow) and the inner $T = 1$ capsid (green), exposed in **b** by removal of the front half of the outer layer. Scale bar, 10 nm. Neighbouring hexamers are 10.2 nm apart, centre-to-centre, at mid-floor, whereas the spacing between hexamers and pentamers is 9.5 nm. Scale bar, 10 nm. **c**, Pseudo-atomic model of the $T = 3$ capsid. An asymmetric unit contains three CA subunits: P1 in pentamers (red), and H1 (blue) and H2 (green) in hexamers. In the $T = 3$ shell, pentamers and hexamers interact by means of the CTD–CTD interface between P1 and H1, whereas adjacent hexamers interact by means of H2. **d**, Interactions between the two layers. NTDs of the $T = 1$ capsid (orange) contact CTDs of H1 subunits in the $T = 3$ capsid (blue). Highlighted in yellow is the loop between helices 4 and 5 in the NTD, and, in cyan, the major homology region in the CTD.

Inter-subunit interactions in capsids

Structures for both domains of RSV CA have been determined[13,14]. The resolution of our $T = 1$ reconstruction was high enough to give unambiguous solutions when the domains were fitted separately into it (Fig. 4a, b). Because there are two CTD structures[13,14], we fitted both and chose the one that correlated best with the density map. In the fit, all of the α-helices map on to well-defined elongated densities (Fig. 2c). The map also contains densities attributable to the single-turn 3_{10} helix (residues 152 to 155) and the β-hairpin at the N terminus of mature CA[14,27]. Next, we generated a pseudo-atomic model of the complete capsid (Supplementary Fig. 1). The NTDs account for its protrusions, whereas the floor layer is composed of CTDs. In a pentamer, the CTD of one subunit underlies the NTD of its neighbour. The C terminus of the NTD and the N terminus of its CTD are 1.6 nm apart, connected by the (unseen) four residues not present in either domain structure (Figs 2c and 4a).

In the $T = 1$ RSV capsid, subunits interact by means of three interfaces (Figs 2b and 4b), as observed previously in planar arrays of HIV CA hexamers[21,28]. Pentameric rings are stabilized by two distinct inter-subunit interactions. In the NTD–NTD interaction, helices 1 and 2 of one NTD are close to helices 1 and 3 of its neighbour (Fig. 5a). In the NTD–CTD interaction, the N terminus of helix 4 in the NTD contacts and is almost perpendicular to helices 8 and 11 in the CTD (Fig. 5c). Also, helix 8 is exposed to the α1–α2 loop, in a way not previously noted (Supplementary Fig. 2). Because the corresponding loop region of HIV CA is the site of a conformational change triggered by the maturation inhibitor CAP-1 (ref. 29), the mechanism for its antiviral activity may be interference with the formation of this interface.

Pentamers interact by means of a CTD–CTD dimerization reaction mediated primarily by the two copies of helix 9 present on

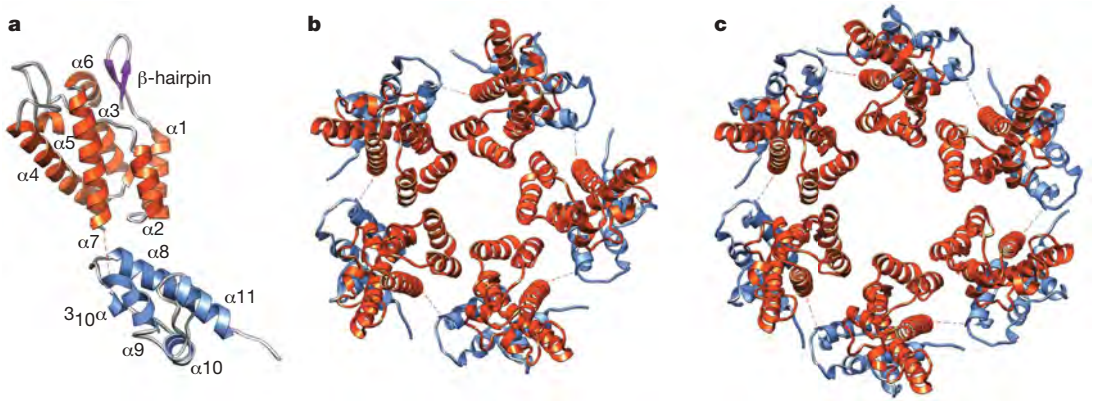


Figure 4 | Pseudo-atomic models of the RSV-CA subunit, pentamer and hexamer. **a**, The relative positions and orientations of the NTD (orange) and CTD (blue), as disposed in the $T = 1$ capsid. The dashed line connects the C terminus of the NTD to the N terminus of the CTD. **b**, **c**, Axial views of the pentamer and hexamer, respectively.

subunits in the two rings (Fig. 5b). These helices cross at an angle of ~45°.

Two hydrophobic patches have been observed on the CTD surface[14], one associated with the C terminus of helix 9 and the other within helices 8 and 11. In the model, they are involved in intermolecular associations at the CTD–CTD and NTD–CTD interfaces, respectively. The patch involved in the NTD–CTD interaction is larger in RSV than in other retroviruses[14].

The $T = 1$ model fitted snugly into the inner shell of the 30-nm particle. Its outer shell has three quasi-equivalent CA subunits: one (P1) forming the pentamers, and two (H1 and H2) forming the hexamers (Fig. 3c, d). After inserting a subunit from the $T = 1$ model at each location, the positions of NTDs and CTDs were refined automatically. In the resulting model (Fig. 3c), pentamers and hexamers interact by means of the CTD–CTD interface between subunits P1 and H1, whereas hexamers interact via the CTD–CTD interface between two H2 subunits. All inter-subunit interactions are very similar to those in the $T = 1$ capsid. In the hexamers (Fig. 4c), NTD–NTD interactions differ slightly in the spacing between neighbours. At the NTD–CTD interface, the angle between helices 4 and 8 ranges between 82° and 87°, whereas the angle between the dimerizing helices 9 at the CTD–CTD interfaces is ~50° (compare to ~45° in the $T = 1$ capsid)—presumably, to accommodate the larger curvature of the $T = 3$ shell.

Electrostatics may control the formation of pentamers

Native RSV capsids typically have 250–300 hexamers but only 12 pentamers[4], and similar numbers pertain for HIV[22,29]. *In vitro* studies have realized many hexamer assemblies[17–19,21], but, until now, no demonstrable pentamers. Thus, hexamers tend to be the favoured oligomer. In contrast, it seems that our assembly conditions boost pentamer production. Two observations indicate that modulation of electrostatic interactions by the high concentration of phosphate ions is a significant factor.

First, CA rings are stabilized by two inter-subunit interactions: NTD–NTD and NTD–CTD. In our models, we see no substantial hydrophobic component at NTD–NTD interfaces, indicating that these interactions are relatively weak. Moreover, in the pentamer and in at least one of the intra-hexamer interfaces, two positively charged residues, Lys 17 in one NTD and Arg 21 in its neighbour, are close together, as are another pair of like charges, Arg 27 and Lys 29 (Supplementary Fig. 2a, b). The electrostatic repulsions encurred by these juxtapositions—which are stronger for the pentamer, in which the residues are closer—must be overcome for the rings to form; it is possible that this is accomplished by a charge-screening effect of the phosphate ions.

Second, the NTD–CTD interaction involves the MHR (Supplementary Fig. 2). Some mutations in the MHR that perturb capsid formation are compensated for by secondary mutations[30] that also map in the NTD–CTD interface[26,31]. One of these suppressors, R185W, whose mutation alone increases the propensity of CA to assemble, affects a cluster of basic residues that are close to the interface (Supplementary Fig. 2c, d). In this context, R185W may suppress by reducing the charge repulsion.

Nesting involves an alternative nucleation mechanism

Multilayered capsids have been observed in virions[2,4,7] and in *in vitro* assembly[26]. Although apparently aberrant, they nevertheless point to the existence of an alternative assembly pathway that is to be avoided in the normal course of events. Multilayers involve the initial formation of one layer, on which subsequent layers are deposited. Thus, their assembly is nucleated by a different, out-of-plane, mechanism. Our 30-nm particles represent a prototypic multilayer. Their nucleating principle is the inner $T = 1$ capsid. Its interaction with the outer layer is mediated by the loop between helices 4 and 5 (residues 87 to 102) of its NTD making contact with the CTD strand (residues 156 to 163) of hexamer subunit H1 in the outer shell (Fig. 3d). Because inner layer pentamers interact with outer layer hexamers, it is likely that the hexamers are the first part of the outer layer to form, with pentamers subsequently filling the gaps.

Comparison with an HIV CA hexamer

Recently, a model of an HIV CA hexamer was derived by electron crystallography[21]. Comparison of this hexamer to our pentamers and hexamers (Supplementary Fig. 3) highlights the conservation of the three inter-subunit interfaces. These structures differ most in the respective dihedral angles between NTD rings, which correlate with their curvatures—high curvatures in the capsids and zero curvature in the sheets. Flexibility at inter-subunit interfaces is greatest at the NTD–NTD contacts (Supplementary Fig. 4). CTD–CTD interfaces involve pairing of helices 9 at slightly different crossing angles. Compliance in this parameter could facilitate the formation of lattices with different curvatures. Most native capsids have variable curvatures intermediate between those of the $T = 3$ shell and of sheets.

The flexible linker and polymorphism

The present results afford strong support for the concept that most HIV capsids are fullerene cones[17,22] and those of RSV are irregular polyhedra[4]. Thus, retroviral capsids are geometrically related to conventional icosahedral capsids but exhibit an unprecedented degree of polymorphism. Because structural properties tend to be selected for functional advantage, the question arises: how could polymorphism promote replication? With the cellular trafficking protein clathrin, which also forms variable fullerene-like lattices (in this case, from flexible trimers[32]), the functional rationale is evident: it allows cargoes of varying size to be accommodated. No such requirement applies to

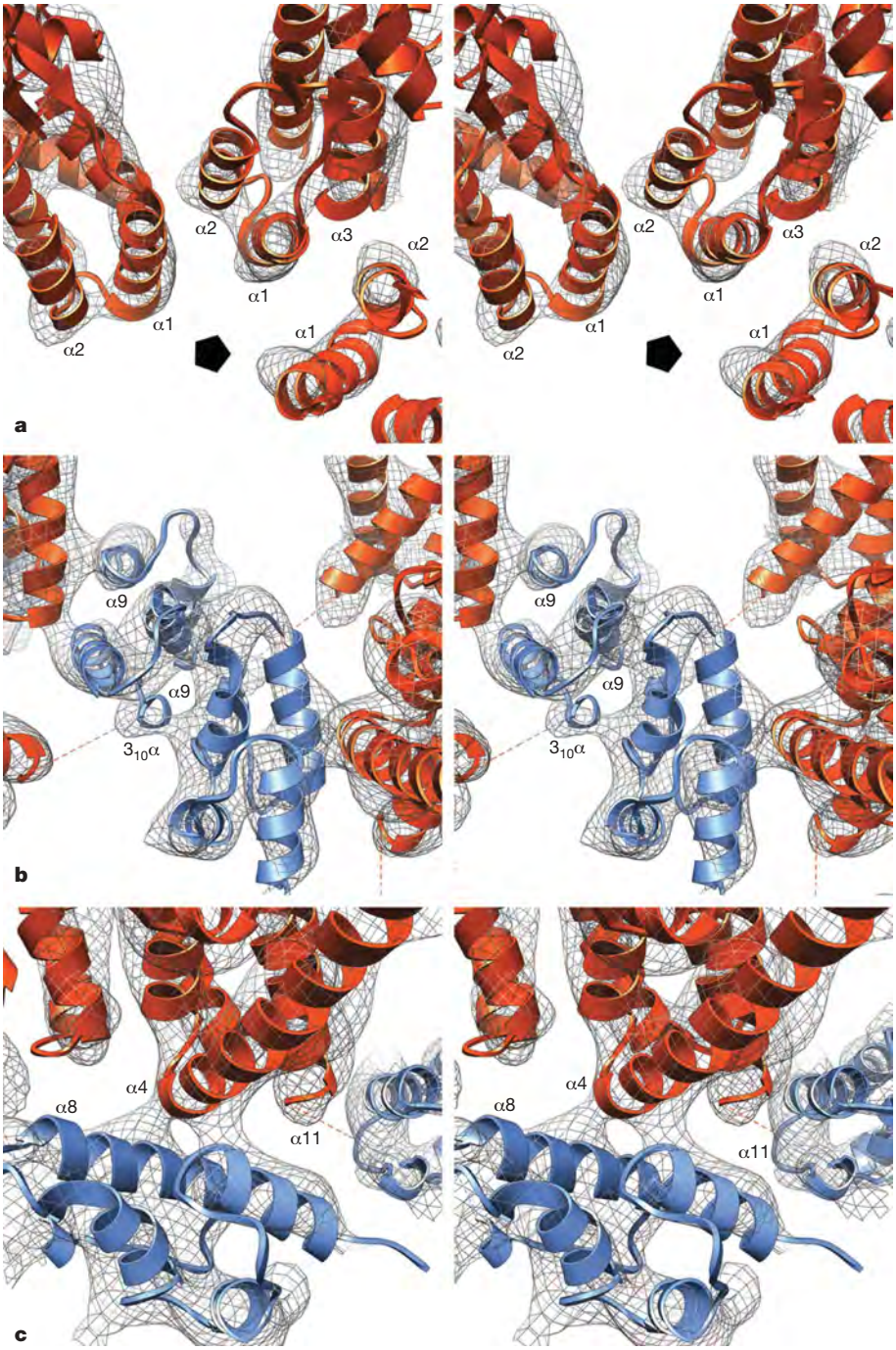


Figure 5 | Inter-subunit interactions in the *T* = 1 capsid, represented in stereo views. Each subunit interacts with three adjacent subunits by means of three interfaces. **a**, NTD–NTD interface. In each pentamer is a ring of five NTDs. The pentagon marks the five-fold axis. Helices 1 and 2 of one subunit are close to helices 1 and 3 of the adjacent subunit. **b**, Neighbouring CA pentamers interact by means of a CTD–CTD dimer interface, mediated by association of helices 9, but presumably also involving the 3_{10} helix. The helices 9 cross at ~45°, their closest point being at the N termini of these helices (residue Ala 184). **c**, NTD–CTD interface. The CTD of one subunit lies under the NTD of the next subunit in the same pentamer. The CTD residues closest to this interface are in the middle of helix 8, around Arg 170, and at the start of helix 11, around Glu 217. The NTD residues involved are at the start of helix 4.

retroviral CAs and no advantage from their polymorphism is yet evident. Be that as it may, the question remains of how they express quasi-equivalence[23]. Here, the answer seems to lie in the flexible linker (Supplementary Fig. 5). Capsids with uniquely defined structures tend to have relatively rigid building blocks and comply with quasi-equivalence through variability at inter-capsomer interfaces. Retroviral capsids, in contrast, have hinged subunits that interact by means of three interfaces, NTD–NTD, CTD–CTD and CTD–NTD, at each of which some play is tolerated, allowing a prolific range of polymorphism.

METHODS SUMMARY

The conditions used for expression of wild-type and I190V mutant RSV CA proteins and for *in vitro* assembly have been described[26]. Focal pairs of vitrified specimens on holey carbon films were recorded on film with a Philips CM200-FEG electron microscope, operating at 120 kV, ×50,000 magnification, and electron doses of ~14 e^- per Å^2 per exposure. Negatives were digitized with a Nikon Super CoolScan 9000 ED at a sampling rate corresponding to 1.27 Å per pixel. Individual particles were selected from the micrographs and preprocessed[33]. An initial three-dimensional model for the 17-nm particle was derived by two-dimensional reference-free classification[34] and icosahedral reconstruction from the class average, corresponding to a three-fold view. An initial model

for the 30-nm particle was generated by a variance-analysis-based procedure[35]. Iterative alignment and reconstructions were performed using PFT2 and EM3DR2 (ref. 36). Density maps were visualized in Chimera[37]. In the case of the 17-nm particle, a pseudo-atomic model for the CA subunit was derived by combining the separate fits of the domains NTD (PDB accession number 1em9)[14] and CTD (PDB accession number 1d1d)[13], which were obtained using an automated rigid-body fitting procedure[38]. A pseudo-atomic model of the 30-nm particle was obtained by an iterative semi-automatic fitting procedure[37], using the subunit solution found for the 17-nm particle as a starting model.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 27 October; accepted 15 December 2008.

1. Vogt, V. M. in *Retroviruses* (eds Coffin, J. M., Hughes, S. H. & Varmus, H.) 27–70 (Cold Spring Harbor Laboratory, 1997).
2. Benjamin, J., Ganser-Pornillos, B. K., Tivol, W. F., Sundquist, W. I. & Jensen, G. J. Three-dimensional structure of HIV-1 virus-like particles by electron cryotomography. *J. Mol. Biol.* **346,** 577–588 (2005).
3. Kingston, R. L., Olson, N. H. & Vogt, V. M. The organization of mature Rous sarcoma virus as studied by cryoelectron microscopy. *J. Struct. Biol.* **136,** 67–80 (2001).
4. Butan, C., Winkler, D. C., Heymann, J. B., Craven, R. C. & Steven, A. C. RSV capsid polymorphism correlates with polymerization efficiency and envelope glycoprotein content: implications that nucleation controls morphogenesis. *J. Mol. Biol.* **376,** 1168–1181 (2008).
5. Yeager, M., Wilson-Kubalek, E. M., Weiner, S. G., Brown, P. O. & Rein, A. Supramolecular organization of immature and mature murine leukemia virus revealed by electron cryo-microscopy: implications for retroviral assembly mechanisms. *Proc. Natl Acad. Sci. USA* **95,** 7299–7304 (1998).
6. Briggs, J. A., Wilk, T., Welker, R., Kräusslich, H. G. & Fuller, S. D. Structural organization of authentic, mature HIV-1 virions and cores. *EMBO J.* **22,** 1707–1715 (2003).
7. Briggs, J. A. *et al.* The mechanism of HIV-1 core assembly: insights from three-dimensional reconstructions of authentic virions. *Structure* **14,** 15–20 (2006).
8. Gamble, T. R. *et al.* Crystal structure of human cyclophilin A bound to the amino-terminal domain of HIV-1 capsid. *Cell* **87,** 1285–1294 (1996).
9. Momany, C. *et al.* Crystal structure of dimeric HIV-1 capsid protein. *Nature Struct. Biol.* **3,** 763–770 (1996).
10. Gitti, R. K. *et al.* Structure of the amino-terminal core domain of the HIV-1 capsid protein. *Science* **273,** 231–235 (1996).
11. Gamble, T. R. *et al.* Structure of the carboxy-terminal dimerization domain of the HIV-1 capsid protein. *Science* **278,** 849–853 (1997).
12. Khorasanizadeh, S., Campos-Olivas, R. & Summers, M. F. Solution structure of the capsid protein from the human T-cell leukemia virus type-I. *J. Mol. Biol.* **291,** 491–505 (1999).
13. Campos-Olivas, R., Newman, J. L. & Summers, M. F. Solution structure and dynamics of the Rous sarcoma virus capsid protein and comparison with capsid proteins of other retroviruses. *J. Mol. Biol.* **296,** 633–649 (2000).
14. Kingston, R. L. *et al.* Structure and self-association of the Rous sarcoma virus capsid protein. *Structure* **8,** 617–628 (2000).
15. Cornilescu, C. C., Bouamr, F., Yao, X., Carter, C. & Tjandra, N. Structural analysis of the N-terminal domain of the human T-cell leukemia virus capsid protein. *J. Mol. Biol.* **306,** 783–797 (2001).
16. Mortuza, G. B. *et al.* High-resolution structure of a retroviral capsid hexameric amino-terminal domain. *Nature* **431,** 481–485 (2004).
17. Li, S., Hill, C. P., Sundquist, W. I. & Finch, J. T. Image reconstructions of helical assemblies of the HIV-1 CA protein. *Nature* **407,** 409–413 (2000).
18. Mayo, K. *et al.* Analysis of Rous sarcoma virus capsid protein variants assembled on lipid monolayers. *J. Mol. Biol.* **316,** 667–678 (2002).
19. Ganser, B. K., Cheng, A., Sundquist, W. I. & Yeager, M. Three-dimensional structure of the M-MuLV CA protein on a lipid monolayer: a general model for retroviral capsid assembly. *EMBO J.* **22,** 2886–2892 (2003).
20. Mayo, K. *et al.* Retrovirus capsid protein assembly arrangements. *J. Mol. Biol.* **325,** 225–237 (2003).
21. Ganser-Pornillos, B. K., Cheng, A. & Yeager, M. Structure of full-length HIV-1 CA: a model for the mature capsid lattice. *Cell* **131,** 70–79 (2007).
22. Ganser, B. K., Li, S., Klishko, V. Y., Finch, J. T. & Sundquist, W. I. Assembly and analysis of conical models for the HIV-1 core. *Science* **283,** 80–83 (1999).
23. Caspar, D. L. & Klug, A. Physical principles in the construction of regular viruses. *Cold Spring Harb. Symp. Quant. Biol.* **27,** 1–24 (1962).
24. Ganser-Pornillos, B. K., von Schwedler, U. K., Stray, K. M., Aiken, C. & Sundquist, W. I. Assembly properties of the human immunodeficiency virus type 1 CA protein. *J. Virol.* **78,** 2545–2552 (2004).
25. Heymann, J. B., Butan, C., Winkler, D. C., Craven, R. C. & Steven, A. C. Irregular and semi-regular polyhedral models for Rous sarcoma virus cores. *Comput. Math. Meth. Medicine* **9,** 197–210 (2008).
26. Purdy, J. G., Flanagan, J. M., Ropson, I. J., Rennoll-Bankert, K. E. & Craven, R. C. Critical role of conserved hydrophobic residues within the major homology region in mature retroviral capsid assembly. *J. Virol.* **82,** 5951–5961 (2008).
27. von Schwedler, U. K. *et al.* Proteolytic refolding of the HIV-1 capsid protein amino-terminus facilitates viral core assembly. *EMBO J.* **17,** 1555–1568 (1998).
28. Lanman, J. *et al.* Identification of novel interactions in HIV-1 capsid protein assembly by high-resolution mass spectrometry. *J. Mol. Biol.* **325,** 759–772 (2003).
29. Briggs, J. A. *et al.* The stoichiometry of Gag protein in HIV-1. *Nature Struct. Mol. Biol.* **11,** 672–675 (2004).
30. Bowzard, J. B., Wills, J. W. & Craven, R. C. Second-site suppressors of Rous sarcoma virus CA mutations: evidence for interdomain interactions. *J. Virol.* **75,** 6850–6856 (2001).
31. Lokhandwala, P. M., Nguyen, T. L., Bowzard, J. B. & Craven, R. C. Cooperative role of the MHR and the CA dimerization helix in the maturation of the functional retrovirus capsid. *Virology* **376,** 191–198 (2008).
32. Edeling, M. A., Smith, C. & Owen, D. Life of a clathrin coat: insights from clathrin and AP structures. *Nature Rev. Mol. Cell Biol.* **7,** 32–44 (2006).
33. Heymann, J. B. & Belnap, D. M. Bsoft: Image processing and molecular modeling for electron microscopy. *J. Struct. Biol.* **157,** 3–18 (2007).
34. Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: Semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128,** 82–97 (1999).
35. Cantele, F., Lanzavecchia, S. & Bellon, P. L. The variance of icosahedral virus models is a key indicator in the structure determination: a model-free reconstruction of viruses, suitable for refractory particles. *J. Struct. Biol.* **141,** 84–92 (2003).
36. Bubeck, D. *et al.* Structure of the poliovirus 135S cell-entry intermediate at 10Å resolution reveals the location of an externalized polypeptide that binds to membranes. *J. Virol.* **79,** 7745–7755 (2005).
37. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
38. Chacon, P. & Wriggers, W. Multi-resolution contour-based fitting of macromolecular structures. *J. Mol. Biol.* **317,** 375–384 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Flanagan for advice on protein purification and analytic methods and access to equipment, R. Meyers for assistance in electron microscopy at the Penn State College of Medicine and B. Heymann for advice on data analysis. This work was supported by the Intramural Research Program of NIAMS and the IATAP Program (A.C.S.), and funding from NIH grant CA100322, the Pennsylvania Department of Health and the Penn State Cancer Institute (R.C.C.).

Author Contributions A.C.S. and R.C.C. designed the project; J.G.P. prepared the capsids with guidance from R.C.C.; N.C. performed the cryo-electron microscopy; G.C. performed the image reconstruction and modelling; and A.C.S. and G.C. wrote the paper with input from the other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.C.S. (stevena@mail.nih.gov) or R.C.C. (rcc6@psu.edu).

METHODS

Cryo-electron microscopy. Most particles assembled from wild-type RSV CA were 17-nm capsids. The 30-nm particles analysed were from the I190V mutant, because earlier experiments appraised by negative staining suggested a higher incidence of larger particles with this mutant[26]. However, they were also rare (~1%) in this preparation. Vitrified specimens on holey carbon films were observed on a CM200-FEG electron microscope (FEI), as described[39]. Focal pairs were recorded on film at ×50,000 magnification, with approximate defocus values of −1.0 μm and −1.5 μm, respectively. Negatives were screened by optical diffraction, and 6 focal pairs of wild-type CA capsids and 32 focal pairs of I190V CA capsids were selected for analysis.

Image processing. Image processing was performed with Bsoft[33], unless otherwise stated. Contrast transfer functions (CTFs) were estimated in bshow. A total of 2,871 17-nm particles and 88 30-nm particles were picked manually and these images were phase-flipped. 967 17-nm particles were aligned and classified using the refine2d python macros in EMAN[34,40]. The images were assigned to 20 classes, among which, two-, three- and five-fold views were identified visually. To generate an initial three-dimensional model, icosahedral symmetry was imposed on the three-fold class average. An initial model for the 30-nm particle was generated by the variance analysis-based procedure implemented in VIVA[35]. The same approach was used to validate the initial model for the 17-nm particle. Iterative refinement was performed using PFT2 and EM3DR2[36] (http://people.chem.byu.edu/belnap/). Full CTF correction was applied in calculating the final reconstruction of the 17-nm particle in EM3DR2. A total of 1,478 (17-nm) and 48 (30-nm) particles were used for the final reconstructions. The handedness of the electron microscopy structures was assigned by comparing with the X-ray structure of the NTD (see below). In terms of FSC coefficients[41], the resolution of the 17-nm particle was 0.98 nm (threshold, 0.3) or 1.04 nm (0.5), and that of the 30-nm particle was 2.06 nm (0.3) or 2.26 nm (0.5).

Visualization and fitting of atomic structures. The electron microscopy density maps were band-limited to resolutions of 0.9 nm (17-nm particle) and 1.9 nm (30-nm particle) for visualization purposes. The contour levels used for surface rendering were set to enclose 100% of expected mass, calculated at 25.5 kDa per CA subunit. For the 17-nm particle, atomic models for the two domains were used in an automated rigid-body fitting procedure. Initially the NTD crystal structure (PDB accession number 1em9) and the NMR structure of the CTD (PDB accession number 1eoq)[14] were fitted separately, using colores in the SITUS package[38] with standard parameters and a target resolution of 1.0 nm. All 60 positions were unambiguously identified, without overlapping. Alternative models for the CTD (residues 152–230) were obtained from the NMR structures of the full CA subunit minus the β-hairpin (PDB accession number 1d1d)[13]. These models were first aligned in Chimera to the solution found for 1eoq, and their orientations and positions were refined against the density map using the SITUS program colacor. The correlation coefficients from the fitting of all models were compared, and the best fit selected. The results of the fitting (Fig. 2c) are shown on a density map of the $T = 1$ capsid for which high-resolution Fourier amplitudes were enhanced by using a pseudo-atomic model as a reference. A pseudo-atomic model for the CA subunit was derived by combining the separate fits of the two domains, and a capsid model obtained by imposing icosahedral symmetry.

A different approach was used to derive a pseudo-atomic model for the 30-nm particle. This fitting was performed with the tool Fit in Map in Chimera, using the $T = 1$ model as reference. This model was fitted intact into the inner layer density. Next, a pentamer from this model was fitted into the outer shell pentamer, and a solution was determined for each of the two subunits (H1 and H2) in the hexamer. This solution was refined through three iterations. After each cycle, a model of the complete capsid was generated by imposing icosahedral symmetry to the solutions found for the three subunits. Then, the positions of the corresponding NTDs and CTDs were refined separately, in the order: H1 NTD, H2 NTD, P1 CTD, H1 CTD and H2 CTD. The P1 NTD was not modified after the initial fitting. For each subunit, the refinement was performed on a density map created by subtracting out the densities associated with all subunits except for the one under refinement (tools Color Zone and Split Map in Chimera).

To obtain a model of the HIV-1 CA hexamer for comparative purposes, we fitted published models of the NTD (PDB accession number 1gwp, residues 1–148)[42] and the CTD (PDB accession number 1a43, residues 149–219)[43] into the density map of two-dimensional crystals of the R18L mutant of HIV-1 CA (EM Databank accession number EMD-1529)[21]. After manually fitting the two domains in Chimera, the solutions were separately refined using the SITUS program colacor and a target resolution of 0.9 nm. This solution agrees with the one derived in ref. 21 (PDB accession number 3dik) to within a root mean squared deviation of 0.2 nm.

39. Cheng, N. *et al.* Handedness of the herpes simplex virus capsid and procapsid. *J. Virol.* **76,** 7855–7859 (2002).
40. Chen, D. H., Song, J. L., Chuang, D. T., Chiu, W. & Ludtke, S. J. An expanded conformation of single-ring GroEL–GroES complex encapsulates an 86 kDa substrate. *Structure* **14,** 1711–1722 (2006).
41. Saxton, W. O. & Baumeister, W. The correlation averaging of a regularly arranged bacterial cell envelope protein. *J. Microsc.* **127,** 127–138 (1982).
42. Tang, C., Ndassa, Y. & Summers, M. F. Structure of the N-terminal 283-residue fragment of the immature HIV-1 Gag polyprotein. *Nature Struct. Biol.* **9,** 537–543 (2002).
43. Worthylake, D. K., Wang, H., Yoo, S., Sundquist, W. I. & Hill, C. P. Structures of the HIV-1 capsid protein dimerization domain at 2.6 Å resolution. *Acta Crystallogr. D* **55,** 85–92 (1999).

LETTERS

A kiloparsec-scale hyper-starburst in a quasar host less than 1 gigayear after the Big Bang

Fabian Walter[1], Dominik Riechers[1,2], Pierre Cox[3], Roberto Neri[3], Chris Carilli[4], Frank Bertoldi[5], Axel Weiss[6] & Roberto Maiolino[7]

The host galaxy of the quasar SDSS J114816.64+525150.3 (at redshift $z = 6.42$, when the Universe was less than a billion years old) has an infrared luminosity of 2.2×10^{13} times that of the Sun[1,2], presumably significantly powered by a massive burst of star formation[3–6]. In local examples of extremely luminous galaxies, such as Arp 220, the burst of star formation is concentrated in a relatively small central region of <100 pc radius[7,8]. It is not known on which scales stars are forming in active galaxies in the early Universe, at a time when they are probably undergoing their initial burst of star formation. We do know that at some early time, structures comparable to the spheroidal bulge of the Milky Way must have formed. Here we report a spatially resolved image of [C II] emission of the host galaxy of J114816.64+525150.3 that demonstrates that its star-forming gas is distributed over a radius of about 750 pc around the centre. The surface density of the star formation rate averaged over this region is $\sim$1,000$M_{\odot}$ year^{-1} kpc^{-2}. This surface density is comparable to the peak in Arp 220, although about two orders of magnitude larger in area. This vigorous star-forming event is likely to give rise to a massive spheroidal component in this system.

The forbidden $^2P_{3/2} \rightarrow {}^2P_{1/2}$ fine-structure line of ionized carbon ([C II]) at 158 μm provides effective cooling in regions where atomic transitions cannot be excited, and therefore helps gas clouds to contract and form stars. [C II] emission is thus known to be a fundamental diagnostic tool of the star-forming interstellar medium[9,10]. Given the very bright continuum emission of the central accreting black hole of quasars in optical and near-infrared wavebands, standard star formation tracers (such as hydrogen recombination lines) cannot be used to study star formation in these systems. The [C II] line, however, is much brighter than the underlying far-infrared (FIR) continuum, thus making it a prime choice to characterize star formation in quasar host galaxies.

We used the IRAM Plateau de Bure interferometer to resolve the [C II] emission from the $z = 6.42$ host galaxy of J114816.64+525150.3 (one of the most distant quasars known[11,12]; hereafter J1148+5251) with a linear resolution of $\sim$1.5 kpc. J1148+5251 is one of only two sources for which [C II] emission has yet been detected at high redshift[5,13]. A large reservoir of molecular gas ($2 \times 10^{10} M_{\odot}$), the prerequisite for star formation, has been characterized in this system through redshifted rotational transition lines of carbon monoxide[3,4,14]. At a redshift of $z = 6.42$, the age of the Universe was just $\sim$870 million years (or 1/16th of its present age) and 1″ on the sky corresponds to 5.6 kpc[15,16].

The distribution of the [C II] emission is shown in Fig. 1b. Gaussian fitting to the spatially resolved [C II] emission gives an intrinsic source size of 0.27″ ± 0.05″ (1.5 ± 0.3 kpc). The [C II] emission is embedded within the molecular gas reservoir traced by CO (ref. 14), but the [C II] emission is offset to the north from the optical quasar and the CO peak by $\sim$0.1″ ($\sim$600 pc). Given the good agreement between the position of the optical quasar and the simultaneous 158-μm continuum observations (Fig. 1a) we do not attribute this offset to inaccurate astrometry. The significance of the [C II] detection is high enough that it shows spatially resolved velocity structure (red and blue contours in Fig. 1c).

The (rest-frame) FIR continuum emission underlying the [C II] line is detected at 10-sigma significance in the integrated frequency spectrum (Fig. 2). If the FIR continuum was due to the (unresolved) optical quasar, a 10-sigma point source is expected at the optical position. However, from Fig. 1a we find only $\sim$50% of the flux to be coincident with the optical quasar position. This implies that the sensitivity of our observations is not high enough to image the remaining FIR flux that is presumably due to the more extended emission from star formation. This would imply that at most 50% of the FIR emission can be attributed to heating by the central black hole; that is, the FIR emission is significantly powered by star formation (in good agreement with the molecular gas[3,4], dense gas[17], radio continuum[6] and dust properties[1,2] of this source). In the following we thus assume a FIR luminosity due to star formation of $\sim 1.1 \times 10^{13} L_{\odot}$, that is, a star formation rate of $\sim$1,700$M_{\odot}$ yr^{-1} (assuming a standard initial stellar mass function[1,18]). The low significance of the resolved FIR emission is the reason that it cannot be used to constrain the size of the starburst region.

The compactness of the [C II] emission implies that massive star formation is concentrated in the central region with radius 750 pc of the system, even though molecular material is available on larger scales (but our [C II] observations cannot rule out star formation at lower surface densities over the entire molecular gas reservoir). Given the star formation rate derived above, we find an extreme average star formation rate surface density of $\sim$1,000$M_{\odot}$ yr^{-1} kpc^{-2} ($\sim 7 \times 10^{12} L_{\odot}$ pc^{-2}) over this central 750-pc radius region. Similarly high starburst surface densities are also found in the centre of local ultraluminous infrared galaxies such as Arp 220 (where each nucleus of size $\sim$100 pc has $L_{\rm FIR} = 3 \times 10^{11} L_{\odot}$), albeit on spatial scales that are two orders of magnitudes smaller[7,8]. For comparison, the young star-forming cluster associated with Orion KL also shows such high densities in its central region[19] ($L_{\rm FIR} = 1.2 \times 10^5 L_{\odot}$, area $\sim$1 arcmin^2, 0.013 pc^2, resulting in $\sim 10^{13} L_{\odot}$ kpc^{-2}), but over an area that is eight orders of magnitudes smaller than in J1148+5251.

In the context of other galaxies in the early Universe, this kiloparsec-scale 'hyper-starburst' has a star formation rate with a surface density an order of magnitude higher than that found in massive star-forming $z \approx 2.5$ submillimetre galaxies[20]. It is, however, consistent

[1]Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany. [2]California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA. [3]Institut de Radio Astronomie Millimétrique, 300 rue de la Piscine, F-38406 St-Martin-d'Hères, France. [4]National Radio Astronomy Observatory, PO Box O, Socorro, New Mexico 87801, USA. [5]Argelander Institut für Astronomie, Auf dem Hügel 71, D-53121 Bonn, Germany. [6]Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, D-53121 Bonn, Germany. [7]L'Istituto Nazionale di Astrofisica, Osservatorio Astronomico di Roma, I-00040 Monte Porzio Catone, Roma, Italy.

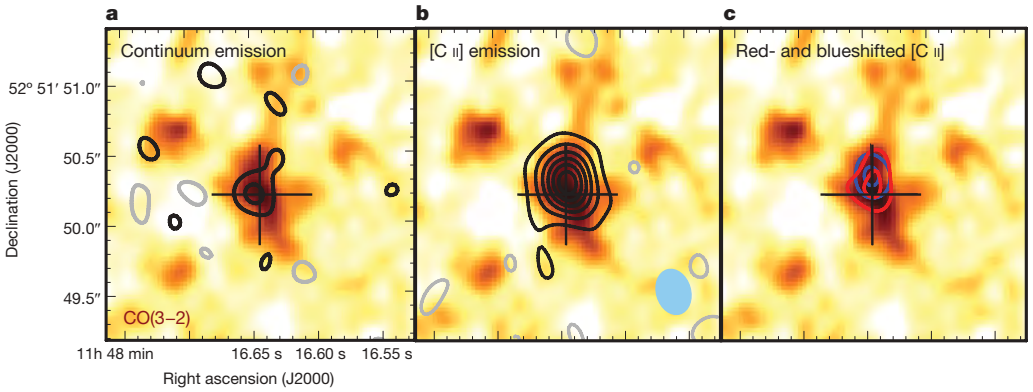


Figure 1 | [C II] observations of the $z = 6.42$ quasar J1148+5251 obtained with the IRAM Plateau de Bure interferometer. Observations were obtained in the most extended antenna configuration during three tracks in early 2007 and 2008 ($\nu_{obs} = 256.17$ GHz, $\nu_{rest} = 1{,}900.54$ GHz), resulting in a resolution of $0.31'' \times 0.23''$ (1.7 kpc × 1.3 kpc; the beam size is shown in light blue in **b**). The resolved CO emission from observations made with the Very Large Array[14] is displayed as a colour scale in all three panels. The cross indicates the absolute position (uncertainty: 0.03″) of the (unresolved) optical quasar as derived from Hubble Space Telescope observations[26]. **a**, Contours represent the far-infrared continuum emission obtained from the line-free channels of the [C II] observations integrated over a bandwidth of 445 km s^{-1} (contour levels are –0.9 (grey), 0.9 and 1.8 mJy (black); r.m.s. noise: 0.45 mJy). There is good agreement between the optical quasar and the peak of the continuum emission, as well as the peak of the molecular gas emission traced by CO, demonstrating that our astrometry is accurate on scales of <0.1″. **b**, Contours show the [C II] emission over a velocity range of –293 to +293 km s^{-1} (contours are plotted in steps of 0.72 mJy; r.m.s. noise: 0.36 mJy). The (rest-frame) beam-averaged peak brightness temperature of the [C II] emission is 9.4 ± 0.9 K (from the peak flux of 7.0 ± 0.36 mJy at a resolution of $0.31'' \times 0.23''$), which is similar to the CO brightness temperature (8.3 K)[14]. If the intrinsic temperature of the gas were similar to that of the dust[2] (30–50 K), this would imply that we have not fully resolved the CO or the [C II] emission. **c**, Contours of blue- and redshifted emission (averaged over velocities from 75–175 km s^{-1} on either side) are plotted as blue and red contours at 3.2 and 4.8 mJy, respectively (r.m.s. noise: 0.63 mJy). The dynamical mass of $\sim 10^{10} M_{\odot}$ within the central 1.5 kpc deduced from these observations (assuming $\nu_{rot} \approx 250$ km s^{-1}) is in agreement with earlier estimates on larger spatial scales[14].

with recent theoretical descriptions of Eddington-limited star formation of a radiation-pressure-supported starburst on kiloparsec scales[21]. The high surface density of the star formation rate is also compatible with other theories describing 'maximum starbursts'[22]: stars can form at a rate limited by SFR = $\varepsilon \times M_{gas}/t_{dyn}$, where ε is the star formation efficiency, M_{gas} is the gas within radius r and t_{dyn} is the dynamical (or free-fall) time, given by $\sqrt{r^3/(2\,GM)}$. For $r = 750$ pc and $M \approx M_{gas} \approx 10^{10} M_{\odot}$, a star formation efficiency of $\varepsilon \approx 0.4$ is required to explain the density of star formation rate that we observe in the case of J1148+5251. Such high efficiencies may be expected given the high fractions of dense gas found in local ultraluminous infrared galaxies[23]. In this calculation, we assume that the stellar initial mass function in this object is not significantly different from what is known locally. Such a high star formation efficiency could be expected if J1148+5251 were to undergo a major merger, where the gas is funnelled rapidly to the central 1.5 kpc. We note, however, that our observations do not provide clear evidence for a merging system and that other mechanisms may be responsible for fuelling the ongoing starburst[24]. We also note that the surface density of star formation rate of $\sim 1{,}000 M_{\odot}$ yr^{-1} kpc^{-2} is a value averaged over roughly the central kiloparsec; this value could be significantly higher on smaller scales, which in turn may violate the theoretical descriptions of 'maximum starbursts'.

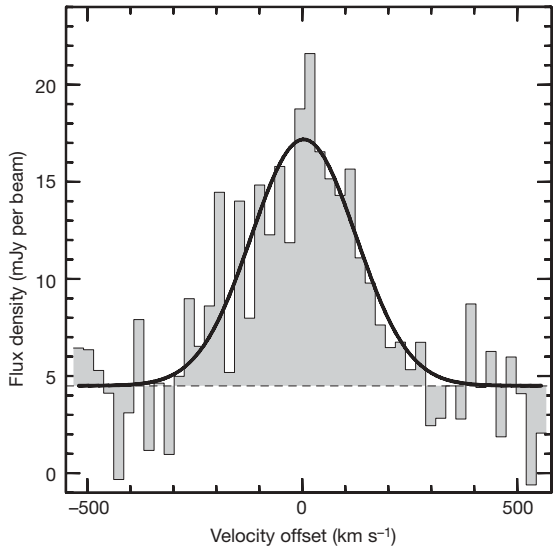


Figure 2 | Spatially integrated [C II] spectrum of the $z = 6.42$ quasar J1148+5251. The [C II] line is detected at high significance (bandwidth covered: 1 GHz, or 1,100 km s^{-1}) and is present on top of a 4.5 ± 0.62 mJy continuum (consistent with an earlier estimate[1] of 5.0 ± 0.6 mJy). Gaussian fitting to the line gives a [C II] peak flux of 12.7 ± 1.05 mJy, a full width at half maximum velocity of 287 ± 28 km s^{-1} and a central velocity of 3 ± 12 km s^{-1} relative to the CO redshift[4] of $z = 6.419$ ($\nu_{obs} = 256.17$ GHz). This leads to a [C II] flux of 3.9 ± 0.3 Jy km s^{-1} (consistent with earlier, unresolved observations[5] of 4.1 ± 0.5 Jy km s^{-1}), which corresponds to a [C II] luminosity[27] of $L'_{[CII]} = 1.90 \pm 0.16 \times 10^{10}$ K km s^{-1} pc^{-2} or $L_{[CII]} = 4.18 \pm 0.35 \times 10^9 L_{\odot}$ (adopting a luminosity distance of $D_L = 64$ Gpc[16]), yielding $L_{[CII]}/L_{FIR} = 1.9 \times 10^{-4}$. This ratio is, by an order of magnitude, smaller than what is found in local star-forming galaxies (a finding consistent with local ultraluminous infrared galaxies[5,28,29]). The line-free channels of the [C II] observations are used to construct a continuum image of J1148+5251 at 158 μm (rest wavelength) as shown in Fig. 1a.

Our observations provide direct evidence for strong, kiloparsec-scale star formation episodes at the end of cosmic reionization that lead to the growth of stellar bulges in quasar host galaxies. Such 'hyper-starbursts' seem to have surface densities of the star formation rate over kiloparsec scales that are an order of magnitude higher than previously studied systems at high redshift[20]. The observations presented here are currently the best means by which to quantify star formation rates and their surface densities in quasars at the earliest cosmic epochs. They thus demonstrate that [C II] observations will have a key role in studies of resolved star formation regions in the first gigayear of the Universe that will be made with the upcoming Atacama Large Millimetre/submillimetre Array (ALMA)[25].

Received 25 July; accepted 18 November 2008.

1. Bertoldi, F. *et al.* Dust and molecular emission from high-redshift quasars. *Astron. Astrophys.* **406**, 55–58 (2003).
2. Beelen, A. *et al.* 350 micron dust emission from high-redshift quasars. *Astrophys. J.* **642**, 694–701 (2006).
3. Walter, F. *et al.* Molecular gas in the host galaxy of a quasar at redshift $z = 6.42$. *Nature* **424**, 406–408 (2003).

4. Bertoldi, F. *et al.* High-excitation CO in a quasar host galaxy at $z = 6.42$. *Astron. Astrophys. Lett.* **409,** 47–50 (2003).
5. Maiolino, R. *et al.* First detection of [C II] 158 μm at high redshift: vigorous star formation in the early universe. *Astron. Astrophys. Lett.* **440,** 51–54 (2005).
6. Carilli, C. *et al.* Radio continuum imaging of far-infrared-luminous QSOs at $z > 6$. *Astron. J.* **128,** 997–1001 (2004).
7. Downes, D. & Solomon, P. Rotating nuclear rings and extreme starbursts in ultraluminous galaxies. *Astrophys. J.* **507,** 615–654 (1999).
8. Scoville, N. Z., Yun, M. S. & Bryant, P. M. Arcsecond imaging of CO emission in the nucleus of Arp 220. *Astrophys. J.* **484,** 702–719 (1997).
9. Tielens, A. G. G. M. & Hollenbach, D. Photodissociation regions. I. Basic model. II. A model for the Orion photodissociation region. *Astrophys. J.* **291,** 722–754 (1985).
10. Stacey, G. J. *et al.* The 158 micron forbidden C II line—a measure of global star formation activity in galaxies. *Astrophys. J.* **373,** 423–444 (1991).
11. Fan, X. *et al.* A survey of $z > 5.7$ quasars in the Sloan Digital Sky Survey. II. Discovery of three additional quasars at $z > 6$. *Astron. J.* **125,** 1649–1659 (2003).
12. Fan, X. *et al.* Constraining the evolution of the ionizing background and the epoch of reionization with $z \sim 6$ quasars. II. A sample of 19 quasars. *Astron. J.* **132,** 117–136 (2006).
13. Iono, D. *et al.* A detection of [C II] line emission in the $z = 4.7$ QSO BR 1202–0725. *Astrophys. J. Lett.* **645,** 97–100 (2006).
14. Walter, F. *et al.* Resolved molecular gas in a quasar host galaxy at redshift z=6.42. *Astrophys. J. Lett.* **615,** 17–20 (2004).
15. Spergel, D. N. *et al.* Three-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Implications for cosmology. *Astrophys. J.* **170** (Suppl.), 377–408 (2007).
16. Wright, E. L. A cosmology calculator for the World Wide Web. *Publ. Astron. Soc. Pacif.* **118,** 1711–1715 (2006).
17. Riechers, D. A., Walter, F., Carilli, C. & Bertoldi, F. observations of dense molecular gas in a quasar host galaxy at z=6.42: Further evidence for a nonlinear dense gas–star formation relation at early cosmic times. *Astrophys. J. Lett.* **671,** 13–16 (2007).
18. Omont, A. *et al.* A 1.2 mm MAMBO/IRAM-30 m survey of dust emission from the highest redshift PSS quasars. *Astron. Astrophys.* **374,** 371–381 (2001).
19. Werner, M. W. *et al.* One arc-minute resolution maps of the Orion Nebula at 20, 50, and 100 microns. *Astrophys. J.* **204,** 420–426 (1976).
20. Tacconi, L. *et al.* High-resolution millimeter imaging of submillimeter galaxies. *Astrophys. J.* **640,** 228–240 (2006).
21. Thompson, T., Quataert, E. & Murrey, N. Radiation pressure-supported starburst disks and active galactic nucleus fueling. *Astrophys. J.* **630,** 167–185 (2005).
22. Elmegreen, B. G. Galactic bulge formation as a maximum intensity starburst. *Astrophys. J.* **517,** 103–107 (1999).
23. Gao, Y. & Solomon, P. M. HCN survey of normal spiral, infrared–luminous, and ultraluminous galaxies. *Astrophys. J.* **152** (Suppl.), 63–80 (2004).
24. Dekel, A. *et al.* Cold streams in early massive hot haloes as the main mode of galaxy formation. *Nature* doi:10.1038/nature07648 (in the press).
25. Walter, F. & Carilli, C. Detecting the most distant ($z > 7$) objects with ALMA. *Astrophys. Space Sci.* **313,** 313–316 (2008).
26. White, R. L., Becker, R. H., Fan, X. & Strauss, M. A. Hubble Space Telescope Advanced Camera for Surveys observations of the $z = 6.42$ quasar SDSS J1148+5251: A leak in the Gunn–Peterson trough. *Astron. J.* **129,** 2102–2107 (2005).
27. Solomon, P. M. & Vanden Bout, P. A. Molecular gas at high redshift. *Annu. Rev. Astron. Astrophys.* **43,** 677–725 (2005).
28. Malhotra, S. *et al.* Infrared Space Observatory measurements of [C II] line variations in galaxies. *Astrophys. J.* **491,** 27–30 (1997).
29. Luhman, M. L. *et al.* Infrared Space Observatory measurements of a [C II] 158 micron line deficit in ultraluminous infrared galaxies. *Astrophys. J. Lett.* **504,** 11–15 (1998).

Acknowledgements This work is based on observations carried out with the IRAM Plateau de Bure Interferometer. IRAM is supported by MPG (Germany), INSU/CNRS (France) and IGN (Spain). D.R. acknowledges support from NASA through a Hubble Fellowship awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA. C.C. acknowledges support from the Max-Planck Gesellschaft and the Alexander von Humboldt Stiftung through the Max-Planck-Forschungspreis 2005. F.W. and D.R. appreciate the hospitality of the Aspen Center for Physics, where this manuscript was written.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.W. (walter@mpia.de).

Vol 457|5 February 2009|doi:10.1038/nature07729

LETTERS

Spin state tomography of optically injected electrons in a semiconductor

Hideo Kosaka[1,2], Takahiro Inagaki[1], Yoshiaki Rikitake[2,3], Hiroshi Imamura[2,4], Yasuyoshi Mitsumori[1,2] & Keiichi Edamatsu[1]

Spin is a fundamental property of electrons, with an important role in information storage[1–4]. For spin-based quantum information technology, preparation and read-out of the electron spin state are essential functions[5–13]. Coherence of the spin state is a manifestation of its quantum nature, so both the preparation and read-out should be spin-coherent. However, the traditional spin measurement technique based on Kerr rotation, which measures spin population using the rotation of the reflected light polarization that is due to the magneto-optical Kerr effect, requires an extra step of spin manipulation or precession to infer the spin coherence[14–20]. Here we describe a technique that generalizes the traditional Kerr rotation approach to enable us to measure the electron spin coherence directly without needing to manipulate the spin dynamics, which allows for a spin projection measurement on an arbitrary set of basis states. Because this technique enables spin state tomography, we call it tomographic Kerr rotation. We demonstrate that the polarization coherence of light is transferred to the spin coherence of electrons, and confirm this by applying the tomographic Kerr rotation method to semiconductor quantum wells with precessing and non-precessing electrons. Spin state transfer and tomography offers a tool for performing basis-independent preparation and read-out of a spin quantum state in a solid.

Electron spin coherence originates in the quantum coherence between up ($\uparrow_e$)/down ($\downarrow_e$) spin states. Similarly, light polarization coherence originates in the quantum coherence between right ($\sigma^+{}_{ph}$)/left ($\sigma^-{}_{ph}$) circular polarization states. It is known that electron spin states can be prepared by injecting circularly polarized photons[5,6,21]. However, we cannot directly prepare a coherent superposition of up and down electron spin states using a conventional preparation scheme. Recently, it was shown that this obstacle can be overcome by using the spin coherence transfer scheme based on the V-shaped band structure in a semiconductor quantum well (Fig. 1a)[22,23]. Spin coherence transfer is essentially different from the coherent control of already existing electron spins based on the optical Stark effect[16], spin-flip Raman scattering[17,24] or Rabi oscillation[18]. An in-plane magnetic field B_x lifts the Kramers degeneracy of the light-hole (subscript LH) spin states and reconfigures them into the superposed $|\pm x\rangle_{LH} = (|\downarrow\rangle_{LH} \pm |\uparrow\rangle_{LH})/\sqrt{2}$ while maintaining the approximate degeneracy of the electron spin states via engineering of the Landé g-factor[15,22]. Spectral selection of one of the light-hole states $|-x\rangle_{LH}$ allows the transition $\alpha|\sigma^+\rangle_{ph} + \beta|\sigma^-\rangle_{ph} \rightarrow (\alpha|\uparrow\rangle_e + \beta|\downarrow\rangle_e) \otimes |-x\rangle_{LH}$, where α and β are complex numbers, and $|\sigma^\pm\rangle_{ph}$ represents a basis vector of the $\sigma^\pm{}_{ph}$ polarization state of light. The electron spin superposition state $\alpha|\uparrow\rangle_e + \beta|\downarrow\rangle_e$ is separated from the hole spin eigenstate $|-x\rangle_{LH}$ and the superposition phase is maintained. Thus spin coherence transfer is advantageous for fast, efficient and reliable preparation of an arbitrary electron spin state.

Reading a coherent spin state is also an important function for spin-based quantum information technology. The electron spin state is conventionally read using the Kerr rotation method, which measures the spin state projection onto one fixed basis defined by the direction of the probe light beam, which is typically normal to the sample surface[14]. Hence, the traditional method of arbitrary spin state projection requires the extra step of manipulating the spin by coherent optical control[19,20]. More conveniently, time-resolved Kerr rotation offers a way of measuring more than one basis projection with the help of spin precession[11]. In contrast, the developed tomographic Kerr rotation (TKR) allows for a spin projection measurement on any arbitrary basis, enabling direct tomographic measurement of the electron spin state or spin coherence without the need to manipulate the spin dynamics. This is because Kerr rotation loses the relative phase information between spin up and down, whereas TKR keeps this phase information, which is about the in-plane spin projections. Therefore the in-plane projection information is lost in Kerr rotation but not lost in TKR. The V-shaped band structure shown in Fig. 1b is again essential here.

We first explain the phenomenological difference between Kerr rotation and TKR by referring to the Poincaré-Bloch spheres shown in Fig. 1c and d. The Kerr rotation (Fig. 1c) measures the rotation of the polarization plane of the linearly polarized probe light upon the reflection from the sample. The Stokes vector of probe light with the $-45°$-tilted linear polarization (D^- or $-y$) state is gyrated around the Bloch vector of the electron spin, then projected along the horizontal/vertical linear polarization (H/V or $\pm x$) basis to give the rotation angle of the Stokes vector around the z-axis, which is proportional to the electron spin polarization projected along the $\uparrow/\downarrow$ ($\pm z$) basis. The rotation of the Stokes vector is restricted to the equator in the Poincaré sphere. But, as shown in Fig. 1d, in the TKR measurement, the rotation of the Stokes vector is not restricted to the equator. For example, the Stokes vector of probe light with the σ^+ polarization ($+z$) state is gyrated around the Bloch vector of the electron spin, and then projected along the horizontal/vertical linear polarization ($\pm x$) basis to give the rotation angle of the Stokes vector around the y-axis, which is proportional to the electron spin polarization projected along the $\uparrow \pm i\downarrow (\pm y)$ basis. The projection basis of the TKR measurement is chosen arbitrarily by the polarization controllers shown in Fig. 1e to perform the spin state tomography.

The rotation of the Stokes vector in the TKR measurement is not restricted to the equator for the following reason. The electron spin state to be measured is a coherent superposition of the electron spin basis states in the V system under an in-plane magnetic field, as shown in Fig. 1b. There exists an exchange interaction $\boldsymbol{s}_1 \cdot \boldsymbol{s}_2$ between the prepared electron spin $\boldsymbol{s}_1$ and the virtually created electron spin $\boldsymbol{s}_2$. Provided that the prepared electron is in $|+y\rangle_e$, the probe light $|P\rangle$ in

[1]Laboratory for Nanoelectronics and Spintronics, Research Institute of Electrical Communication, Tohoku University, Sendai 980-8577, Japan. [2]CREST-JST, Saitama 332-0012, Japan. [3]Department of Information Engineering, Sendai National College of Technology, Sendai 989-3128, Japan. [4]Nanotechnology Research Institute, AIST, Tsukuba 305-8568, Japan.

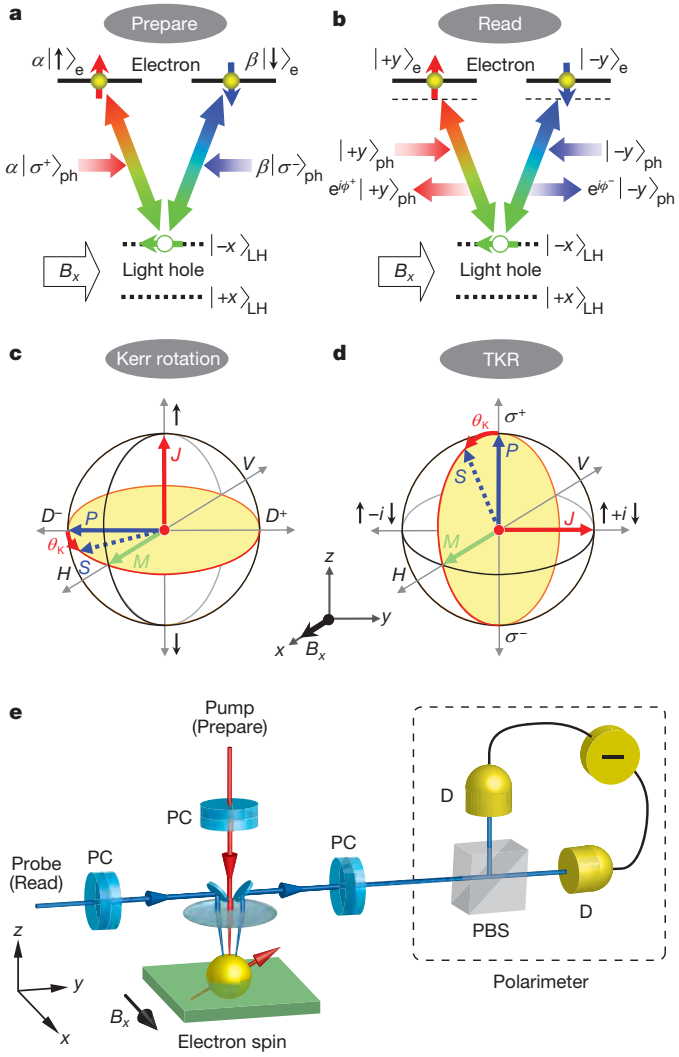


Figure 1 | Operating principle of spin coherence transfer and tomography. **a**, **b**, Three-level V system, consisting of two degenerate electron spin states and one non-degenerate light-hole spin state, enabling the spin coherence transfer ('Prepare') (**a**) and spin state tomography ('Read') (**b**). **c**, **d**, Poincaré-Bloch spheres demonstrating the operating principle of the traditional Kerr rotation (**c**) and the developed TKR (**d**) in their typical configurations. The scattered light (S) rotated from the probe light (P) is projected onto the polarization measurement basis (M), leading to the projection of the electron spin Bloch vector onto the projection basis (J). **e**, Experimental set-up for the TKR. PC, polarization controller; PBS, polarization beam splitter; D, detector. The PCs are composed of a half-wave plate and a quarter-wave plate. The polarimeter measures the light polarization as the difference between the intensities of two orthogonal polarizations.

$|+z\rangle_{ph} = (|+y\rangle_{ph} + |-y\rangle_{ph})/\sqrt{2}$ virtually creates another electron in $|+z\rangle_e = (|+y\rangle_e + |-y\rangle_e)/\sqrt{2}$, together with a light hole in $|-x\rangle_{LH}$. We note that it is essential to choose one of the hole eigenstates. Then, two components $|\pm y\rangle_{ph}$ in the scattered light $|S\rangle$ experience different phase shifts $\phi^{\pm}$ owing to the exchange interaction, that is, $|S\rangle = (e^{i\phi^+}|+y\rangle_{ph} + e^{i\phi^-}|-y\rangle_{ph})/\sqrt{2}$. The Stokes vector thus rotates along the meridian by the Kerr rotation angle θ_K, which is approximately proportional to the phase difference $\phi^+ - \phi^-$ measured by homodyne detection of the scattered light in the measurement basis $|M\rangle = |\pm x\rangle_{ph}$. The θ_K in effect measures the projection of the electron spin onto the projection basis $|J\rangle = |\pm y\rangle_e$. The basis orthogonality among $|P\rangle$, $|M\rangle$ and $|J\rangle$ in the Poincaré sphere is guaranteed by subtracting the TKR amplitude with the opposite probe polarizations $|\pm P\rangle$. The subtraction of the TKR amplitude also guarantees the extraction of only the real part of the optical susceptibility and the elimination of the ambiguity of the complex phase caused by the background reflection from the sample. If necessary, we can easily perform spin state tomography using the Faraday rotation geometry by changing the reflection configuration to a transmission configuration. The theoretical background of this procedure is explained in the Supplementary Information.

Next, we show the results of spin state tomography of two different samples. Sample A is the 11-nm-thick undoped GaAs quantum well embedded in undoped $Al_{0.3}Ga_{0.7}As$. Sample B is the 6-nm-thick undoped GaAs quantum well embedded in undoped $Al_{0.35}Ga_{0.65}As$. The electron and light-hole g-factors under an in-plane magnetic field are respectively -0.21 and -3.5 in sample A, and 0.01 and -2.1 in sample B. A magnetic field $B_x = 7$ T is applied parallel to the quantum well. Because of the difference in the electron g-factors, the electron spins in sample A precess around B_x, whereas no precession is observed in sample B. The details of the experimental set-up are given in the Methods section.

Figure 2 shows the results of the spin state tomography of the precessing electrons in sample A. The fact that the TKR amplitudes for the D^+ (or D^-) probe are delayed by $\pi/2$ from those for σ^+ (or σ^-) indicates that the projection bases are orthogonal to each other in the Bloch sphere. Figure 2b shows the trajectory of the spin Bloch vector reconstructed from the TKR amplitudes shown in Fig. 2a**.** The spin precession starts from the $|+y\rangle_e$ state corresponding to the pump light polarization D^+. The TKR amplitudes for $\sigma^{\pm}$ are smaller than those for $D^{\pm}$ because those for $\sigma^{\pm}$ (or $D^{\pm}$) are proportional to the difference (or sum) of the Kerr spectra originating from the $|+x\rangle_{LH}$ and $|-x\rangle_{LH}$ states with finite spectral widths (see Supplementary Information). Assuming that the norm of the spin vector does not change in time (except for the decoherence effect), we estimate the ratio of the spin projection onto the y axis and that onto the z axis r_{yz} to be 0.34. Figure 2c shows a density matrix of the electron spin state reconstructed from the measured TKR amplitudes at 50 ps (after 2π spin rotation) with the r_{yz} calibration (red and blue bars) together with that of the pump light polarization state in D^+ as a reference (white bars). We assumed that the initial degree of polarization of the electron spin created by a circularly polarized light is unity, owing to the optical selection rule of the well-defined light-hole excitons. This assumption also applies to the D^+ polarized light, because we tuned the pump wavelength such that the yields of the spin transfer for the $\sigma^{\pm}$ and $D^{\pm}$ pump lights are the same[23]. By comparing those two density matrices (blue and white bars), we estimate that the fidelity of the spin coherence transfer from photons to electrons in the maximally superposed case is 0.86 ± 0.03, which is close to unity. The fidelity degradation is due to spin decoherence after the initial preparation[15].

It should be noted that the TKR measurements using heavy-hole states provided negligible amplitudes for the $\sigma^{\pm}$ probe ($\pm y$ projection) compared to those for the $D^{\pm}$ probe ($\pm z$ projection). This means that TKR on the heavy-hole states is effectively just traditional Kerr rotation, because the in-plane g-factor of the heavy-hole state is almost zero owing to the symmetry of the wavefunctions, and thus the hole spin states are degenerate[25].

Figure 3 shows the results of the spin state tomography of the non-precessing electrons in sample B, where the electron g-factor is almost zero. The traditional time-resolved Kerr rotation method cannot be used to measure the spin phase when the precession period is much longer than the decoherence time. In contrast, TKR allows for the spin state tomography of non-precessing electrons. Figure 3a shows the time-resolved TKR amplitudes of the electron spin in the $|+y\rangle$ state projected along the $\pm y$ basis, where no clear spin precession is observed. The TKR amplitudes for the different pump polarizations measured at $\Delta t = 10$ ps (to avoid any contribution from coherent artefacts) are plotted in Fig. 3b**.** From the amplitude difference we estimate the projection ratio r_{yz} to be 0.28, assuming that the yields of the spin transfer for the $\sigma^{\pm}$ and $D^{\pm}$ pump lights are the same. The TKR amplitudes after the calibration as a function of both the pump polarization and the probe polarization (given as the projection

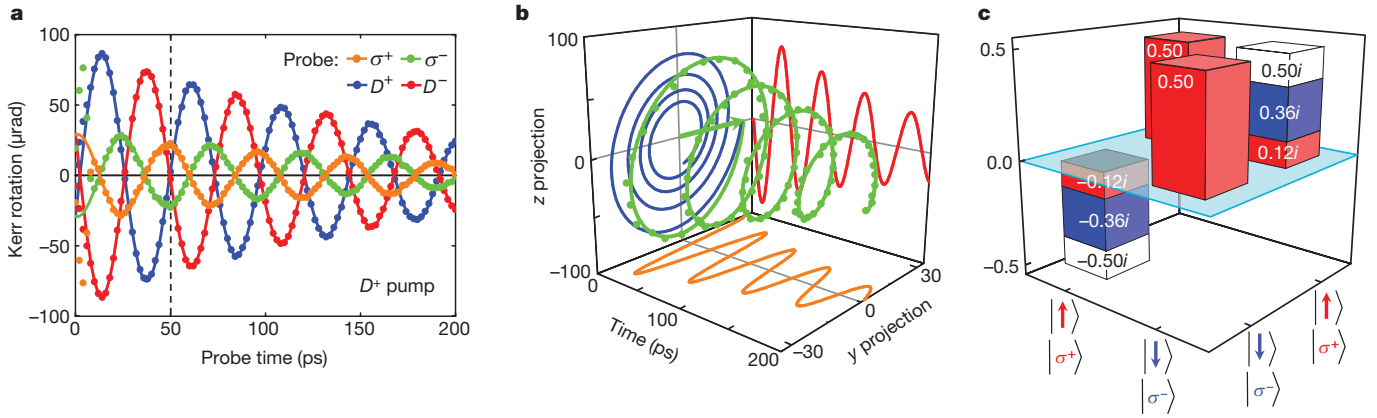


Figure 2 | Spin state tomogram of precessing electrons. **a**, Time-resolved TKR amplitudes pumped with D^+ ($+y$ state) light and probed with σ^+ ($|+y\rangle$ projection, orange), D^+ ($|-z\rangle$ projection, blue), σ^- ($|-y\rangle$ projection, green), or D^- ($|+z\rangle$ projection, red) light with the horizontal/vertical linear polarization ($\pm x$) measurement basis using sample A at $B_x = 7$ T. **b**, A trajectory of the reconstructed electron spin state Bloch vector (green solid line), showing coherent rotation of the electron spin around B_x, starting from the $|+y\rangle$ state (green arrow) corresponding to the D^+ pump polarization. Red, orange and blue solid lines show the projected trajectories. **c**, A density matrix of the optically injected electron spin state in the $\uparrow/\downarrow$ basis reconstructed from the experimental data with the D^+ pump probed after 50 ps (red bars). The TKR amplitudes were calibrated using the y-z projection ratio $r_{yz} = 0.34$ (blue bars). White bars are the density matrix of the pump light in the σ^+/σ^- basis measured by a polarimeter. The vertical axis shows probability amplitude.

state) are shown in Fig. 3c. We confirm the almost ideal phase transfer function $\cos(\phi_x^{ph} - \phi_x^e)$, where ϕ_x^{ph} and ϕ_x^e are phases of the pump light and created electrons in the y-z plane of the Poincaré-Bloch sphere. The polar plot of the calibrated TKR amplitudes shown in Fig. 3d (dots) agrees well with that of the pump light polarization (circles) measured directly by the polarimeter shown in Fig. 1e. The agreement also indicates that quantum coherence is preserved during the spin coherence transfer, even for the system with degenerate electron spin states, in which the traditional Kerr rotation method is not applicable.

The spin coherence transfer and the spin state tomography we have shown here will be applicable to the transfer of a single-particle quantum state and a two-particle entangled state, which is the kind of transfer needed for quantum information technology. Unlike the

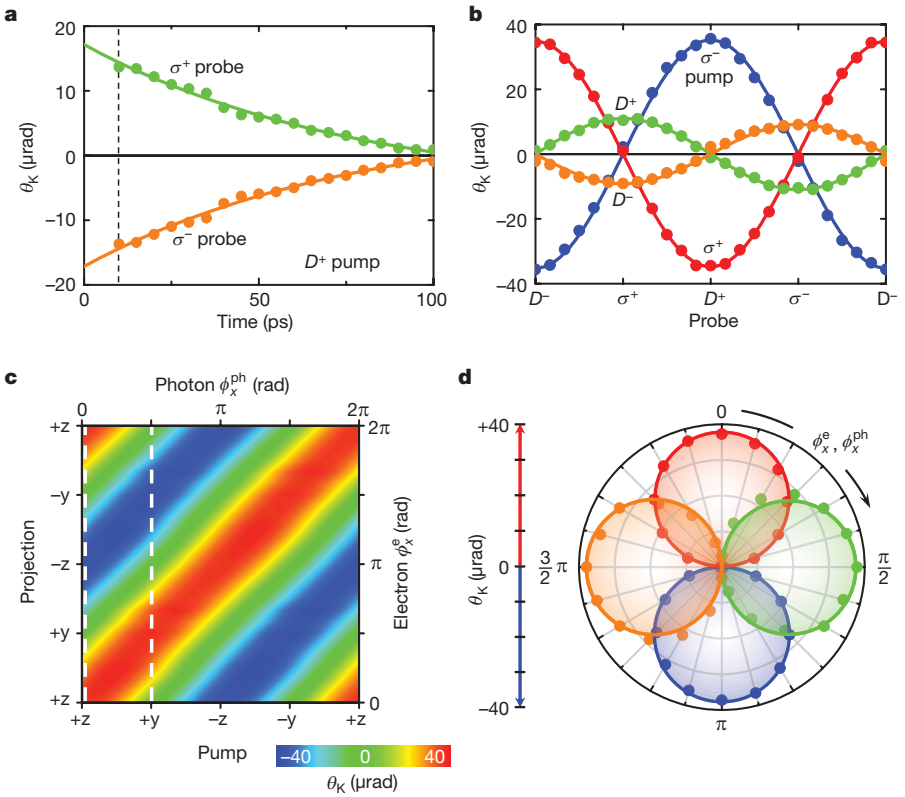


Figure 3 | Spin coherence transfer in the spin degenerate case. **a**, Time-resolved TKR amplitudes (θ_K) with the D^+ pump and $\sigma^\pm$ probe light at $B_x = 7$ T measured using sample B. The curves show the fit to the exponential functions. **b**, Probe light polarization dependence of θ_K with different pump light polarizations measured at $\Delta t = 10$ ps. **c**, θ_K as a function of the pump light states and the electron spin projection states. Calibrated using the estimated y-z projection ratio $r_{yz} = 0.28$ (see text). **d**, Polar plots of the calibrated θ_K (dots) pumped with σ^+ (+z, red and blue) and D^+ (+y, green and orange) light (broken lines in Fig. 3c) as a function of ϕ_x^e. Circles show the polar plot of the pump light polarizations as a function of ϕ_x^{ph}. Red and green (or blue and orange) represent positive (or negative) signs.

electron spin in the quantum well, which does not act as a qubit because it is not individually addressable, the electron spin in a quantum dot does act as a qubit based on the same mechanism. In the case of quantum dots, the light-hole component can hybridize with the heavy-hole state to create superposed eigenstates that can be split even under an in-plane magnetic field. If the splitting is so large that we can spectrally select one of the split states, we can use the state for spin coherence transfer and tomography with the help of appropriate linear transformation. In contrast, in the case of a quantum well, the effect of heavy-hole, light-hole mixing is negligibly small owing to the optical transition at the Γ point. Spin coherence transfer and spin state tomography will provide a way to write and read a spin qubit in a solid-state device[1–4] for quantum cryptography and distributed quantum computing.

METHODS SUMMARY

The pump/probe light bandwidth was set to 0.38 nm (0.74 meV) to configure the V-shaped three-level system shown in Fig. 1a and b at $B_x = 7$ T. Pump wavelengths for the experiments using samples A and B were set to 796.7 nm and 768.4 nm, respectively, which correspond to $|-x\rangle_{LH}$, and the probe wavelengths were set to 795.0 nm and 767.6 nm, respectively which correspond to $|+x\rangle_{LH}$. We used different light-hole states for the pump and probe to avoid their interference effect. The pump and probe lights were generated by spectrally filtering the same light beam from a mode-locked Ti:sapphire laser delivering 130-fs pulses at a repetition rate of 76 MHz, and a variable delay line was set only in the probe path. All the measurements were made at 10 K. Incident angles of the pump and probe light were fixed to nearly normal to the sample surface in the Voigt geometry. This configuration makes TKR essentially different from the longitudinal magneto-optical Kerr effect, which requires tilting of the probe light from the surface normal. The light polarizations of pump, probe and measurement were controlled in any of $\{H, V, D^+, D^-, \sigma^+, \sigma^-\}$ by inserting a pair of half- and quarter-wave plates (Fig. 1e). The influence of the dynamic nuclear-spin polarization was suppressed by periodically alternating the polarization of the pump light by using a photoelastic modulator at a frequency of 42 kHz (ref. 11), which also enables lock-in detection to reveal only the pump-light-induced effect. The 0.5-mW pump light and 0.1-mW probe light for sample A (0.2-mW probe light for sample B) were focused onto a ~100-μm spot, where the average exciton spacing is expected to be much greater than the exciton Bohr radius, thereby eliminating any unnecessary nonlinear effect between excitons.

Received 18 June; accepted 9 December 2008.

1. Loss, D. & DiVincenzo, D. P. Quantum computation with quantum dots. *Phys. Rev. A* **57,** 120–126 (1998).
2. Imamoglu, A. *et al.* Quantum information processing using quantum dot spins and cavity QED. *Phys. Rev. Lett.* **83,** 4204–4207 (1999).
3. Awschalom, D. D., Loss, D. & Samarth, N. (eds) *Semiconductor Spintronics and Quantum Computation* (Springer, 2002).
4. Hanson, R., Kouwenhoven, L. P., Petta, J. R., Tarucha, S. & Vandersypen, L. M. K. Spins in few-electron quantum dots. *Rev. Mod. Phys.* **79,** 1217–1265 (2007).
5. Kroutvar, M. *et al.* Optically programmable electron spin memory using semiconductor quantum dots. *Nature* **432,** 81–84 (2004).
6. Young, R. *et al.* Single electron-spin memory with a semiconductor quantum dot. *New J. Phys.* **9,** 365 (2007).
7. Kikkawa, J. M. & Awschalom, D. D. Lateral drag of spin coherence in gallium arsenide. *Nature* **397,** 139–141 (1999).
8. Bracker, A. S. *et al.* Optical pumping of the electronic and nuclear spin of single charge-tunable quantum dots. *Phys. Rev. Lett.* **94,** 047402 (2005).
9. Berezovsky, J. *et al.* Nondestructive optical measurements of a single electron spin in a quantum dot. *Science* **314,** 1916–1920 (2006).
10. Atature, M., Dreiser, J., Badolato, A. & Imamoglu, A. Observation of Faraday rotation from a single confined spin. *Nature Phys.* **3,** 101–105 (2007).
11. Mikkelsen, M. H. *et al.* Optically detected coherent spin dynamics of a single electron in a quantum dot. *Nature Phys.* **3,** 770–773 (2007).
12. Kosaka, H. *et al.* Single photoelectron trapping, storage, and detection in a field effect transistor. *Phys. Rev. B* **67,** 045104 (2003).
13. Kosaka, H., Mitsumori, Y., Rikitake, Y. & Imamura, H. Polarization transfer from photon to electron spin in g factor engineered quantum wells. *Appl. Phys. Lett.* **90,** 113511 (2007).
14. Baumberg, J. J., Awschalom, D. D., Samarth, N., Luo, H. & Furdyna, J. K. Spin beats and dynamical magnetization in quantum structures. *Phys. Rev. Lett.* **72,** 717–720 (1994).
15. Salis, G. *et al.* Electrical control of spin coherence in semiconductor nanostructures. *Nature* **414,** 619–622 (2001).
16. Gupta, J. A., Knobel, R., Samarth, N. & Awschalom, D. D. Ultrafast manipulation of electron spin coherence. *Science* **292,** 2458–2461 (2001).
17. Dutt, M. V. G. *et al.* Stimulated and spontaneous optical generation of electron spin coherence in charged GaAs quantum dots. *Phys. Rev. Lett.* **94,** 227403 (2005).
18. Greilich, A. *et al.* Optical control of spin coherence in singly charged (In,Ga)As/GaAs quantum dots. *Phys. Rev. Lett.* **96,** 227401 (2006).
19. Wu, Y. *et al.* Density matrix tomography through sequential coherent optical rotations of an exciton qubit in a single quantum dot. *Phys. Rev. Lett.* **96,** 087402 (2006).
20. Wu, Y. *et al.* Selective optical control of electron spin coherence in singly charged GaAs-$Al_{0.3}Ga_{0.7}As$ quantum dots. *Phys. Rev. Lett.* **96,** 097402 (2007).
21. Meier, F. & Zakharchenya, B. P. (eds) *Optical Orientation* Ch. 2 (Elsevier, 1984).
22. Vrijen, R. & Yablonovitch, E. A spin-coherent semiconductor photodetector for quantum communication. *Physica E* **10,** 569–575 (2001).
23. Kosaka, H. *et al.* Coherent transfer of light polarization to electron spins in a semiconductor. *Phys. Rev. Lett.* **100,** 096602 (2008).
24. Xu, X. *et al.* Fast spin state initialization in a singly charged InAs-GaAs quantum dot by optical cooling. *Phys. Rev. Lett.* **99,** 097401 (2007).
25. Marie, X. *et al.* Hole spin quantum beats in quantum-well structures. *Phys. Rev. B* **60,** 5811–5817 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported in part by the Strategic Information and Communications R & D Promotion Program (SCOPE No. 41402001) of the Ministry of Internal Affairs and Communications in Japan.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to H.K. (kosaka@riec.tohoku.ac.jp).

LETTERS

Large-scale pattern growth of graphene films for stretchable transparent electrodes

Keun Soo Kim[1,3,4], Yue Zhao[7], Houk Jang[2], Sang Yoon Lee[5], Jong Min Kim[5], Kwang S. Kim[6], Jong-Hyun Ahn[2,3], Philip Kim[3,7], Jae-Young Choi[5] & Byung Hee Hong[1,3,4]

Problems associated with large-scale pattern growth of graphene constitute one of the main obstacles to using this material in device applications[1]. Recently, macroscopic-scale graphene films were prepared by two-dimensional assembly of graphene sheets chemically derived from graphite crystals and graphene oxides[2,3]. However, the sheet resistance of these films was found to be much larger than theoretically expected values. Here we report the direct synthesis of large-scale graphene films using chemical vapour deposition on thin nickel layers, and present two different methods of patterning the films and transferring them to arbitrary substrates. The transferred graphene films show very low sheet resistance of ~280 Ω per square, with ~80 per cent optical transparency. At low temperatures, the monolayers transferred to silicon dioxide substrates show electron mobility greater than 3,700 $cm^2 V^{-1} s^{-1}$ and exhibit the half-integer quantum Hall effect[4,5], implying that the quality of graphene grown by chemical vapour deposition is as high as mechanically cleaved graphene[6]. Employing the outstanding mechanical properties of graphene[7], we also demonstrate the macroscopic use of these highly conducting and transparent electrodes in flexible, stretchable, foldable electronics[8,9].

Graphene has been attracting much attention owing to its fascinating physical properties such as quantum electronic transport[4,5], a tunable band gap[10], extremely high mobility[11], high elasticity[7] and electromechanical modulation[12]. Since the discovery of the first isolated graphene prepared by mechanical exfoliation of graphite crystals[6], many chemical approaches to synthesize large-scale graphene have been developed, including epitaxial growth on silicon carbide (refs 13, 14) and ruthenium (ref. 15) as well as two-dimensional assembly of reduced graphene oxides[3,16–18] and exfoliated graphene sheets[2]. Epitaxial growth provides high-quality multilayer graphene samples interacting strongly with their substrates, but electrically isolated mono- or bilayer graphene for device applications has not been made. On the other hand, the self-assembly of soluble graphene sheets demonstrates the possibility of low-cost synthesis and the fabrication of large-scale transparent films. However, these assembled graphene films show relatively poor electrical conductivity owing to the poor interlayer junction contact resistance and the structural defects formed during the vigorous exfoliation and reduction processes. In this work, we develop a technique for growing few-layer graphene films using chemical vapour deposition (CVD) and successfully transferring the films to arbitrary substrates without intense mechanical and chemical treatments, to preserve the high crystalline quality of the graphene samples. Therefore, we expect to observe enhanced electrical and mechanical properties. The growth, etching and transferring processes of the CVD-grown large-scale graphene films are summarized in Fig. 1.

It has been known for over 40 years that CVD of hydrocarbons on reactive nickel or transition-metal-carbide surfaces can produce thin graphitic layers[19–21]. However, the large amount of carbon sources absorbed on nickel foils usually form thick graphite crystals rather than graphene films (Fig. 2a). To solve this problem, thin layers of nickel of thickness less than 300 nm were deposited on SiO_2/Si substrates using an electron-beam evaporator, and the samples were then heated to 1,000 °C inside a quartz tube under an argon atmosphere. After flowing reaction gas mixtures (CH_4:H_2:Ar = 50:65:200 standard cubic centimetres per minute), we rapidly cooled the samples to room temperature (~25 °C) at the rate of ~10 °C s^{-1} using flowing argon. We found that this fast cooling rate is critical in suppressing formation of multiple layers and for separating graphene layers efficiently from the substrate in the later process[20].

A scanning electron microscope (SEM; JSM6490, Jeol) image of graphene films on a thin nickel substrate shows clear contrast between areas with different numbers of graphene layers (Fig. 2a). Transmission electron microscope (TEM; JEM3010, Jeol) images (Fig. 2b) show that the film mostly consists of less than a few layers of graphene. After transfer of the film to a silicon substrate with a 300-nm-thick SiO_2 layer, optical and confocal scanning Raman microscope (CRM 200, Witech) images were made of the same area (Fig. 2c, d)[22]. The brightest area in Fig. 2d corresponds to monolayers, and the darkest area is composed of more than ten layers of graphene. Bilayer structures appear to predominate in both TEM and Raman images for this particular sample, which was prepared from 7 min of growth on a 300-nm-thick nickel layer. We found that the average number of graphene layers, the domain size and the substrate coverage can be controlled by changing the nickel thickness and growth time during the growth process (Supplementary Figs 1 and 2), thus providing a way of controlling the growth of graphene for different applications.

Atomic force microscope (AFM; Nanoscopes IIIa and E, Digital Instruments) images often show the ripple structures caused by the difference between the thermal expansion coefficients of nickel and graphene (Fig. 2c, inset; see also Supplementary Fig. 3)[19]. We believe that these ripples make the graphene films more stable against mechanical stretching[23], making the films more expandable, as we will discuss later. Multilayer graphene samples are preferable in terms of mechanical strength for supporting large-area film structures, whereas thinner graphene films have higher optical transparency. We find that a ~300-nm-thick nickel layer on a silicon wafer is the optimal substrate for the large-scale CVD growth that yields mechanically stable, transparent graphene films to be transferred and stretched after they are formed, and that thinner nickel layers with a shorter growth time yield predominantly mono- and bilayer graphene film for microelectronic device applications (Supplementary Fig. 1c).

[1]Department of Chemistry, [2]School of Advanced Materials Science and Engineering, [3]SKKU Advanced Institute of Nanotechnology, [4]Center for Nanotubes and Nanostructured Composites, Sungkyunkwan University, Suwon 440-746, Korea. [5]Samsung Advanced Institute of Technology, PO Box 111, Suwon 440-600, Korea. [6]Department of Chemistry, Pohang University of Science and Technology, Pohang 790-784, Korea. [7]Department of Physics, Columbia University, New York, New York 10027, USA.

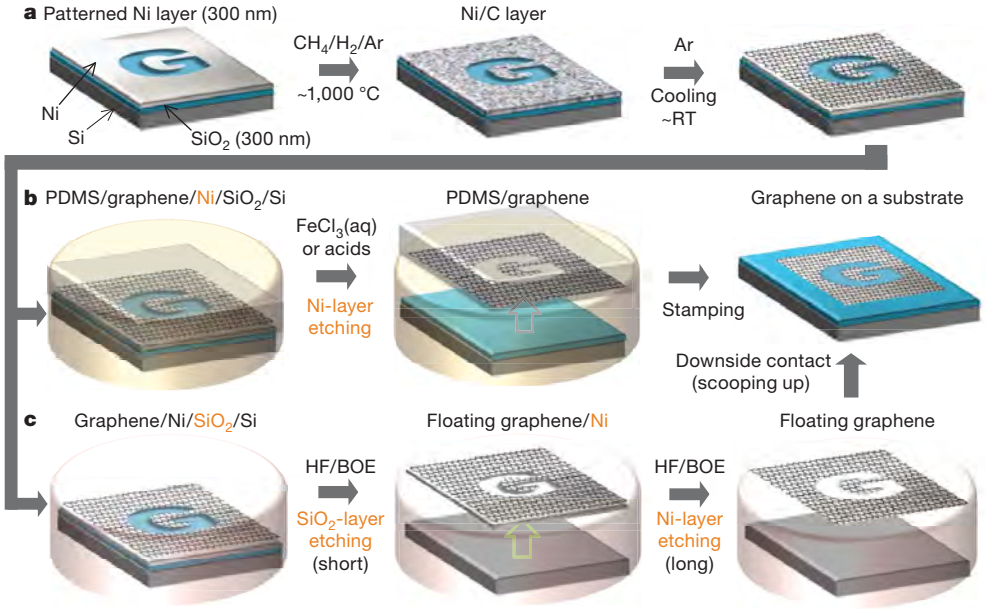


Figure 1 | Synthesis, etching and transfer processes for the large-scale and patterned graphene films. **a**, Synthesis of patterned graphene films on thin nickel layers. **b**, Etching using $FeCl_3$ (or acids) and transfer of graphene films using a PDMS stamp. **c**, Etching using BOE or hydrogen fluoride (HF) solution and transfer of graphene films. RT, room temperature (~25 °C).

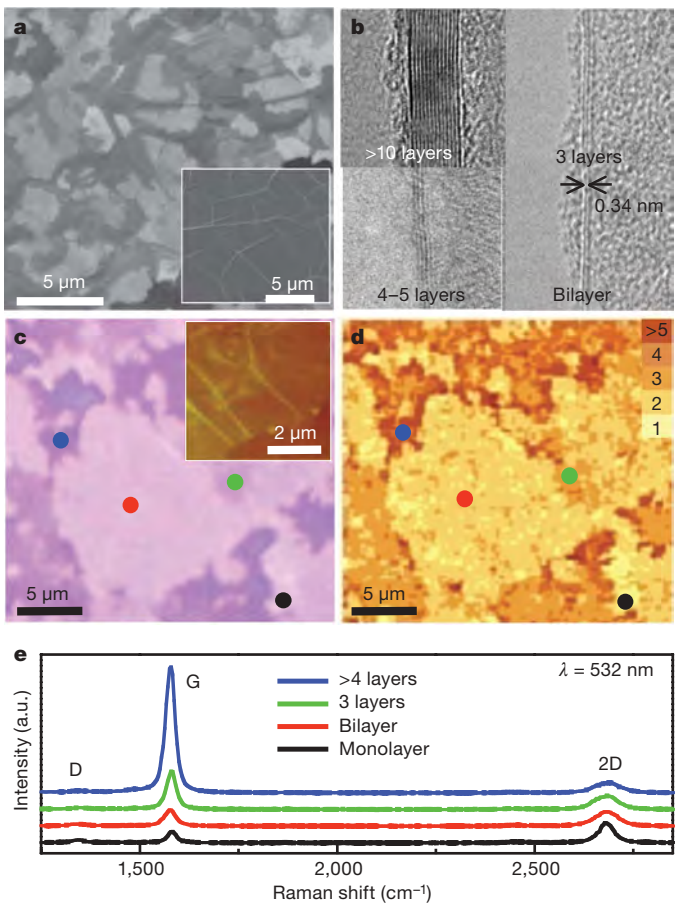


Figure 2 | Various spectroscopic analyses of the large-scale graphene films grown by CVD. **a**, SEM images of as-grown graphene films on thin (300-nm) nickel layers and thick (1-mm) Ni foils (inset). **b**, TEM images of graphene films of different thicknesses. **c**, An optical microscope image of the graphene film transferred to a 300-nm-thick silicon dioxide layer. The inset AFM image shows typical rippled structures. **d**, A confocal scanning Raman image corresponding to **c**. The number of layers is estimated from the intensities, shapes and positions of the G-band and 2D-band peaks. **e**, Raman spectra (532-nm laser wavelength) obtained from the corresponding coloured spots in **c** and **d**. a.u., arbitrary units.

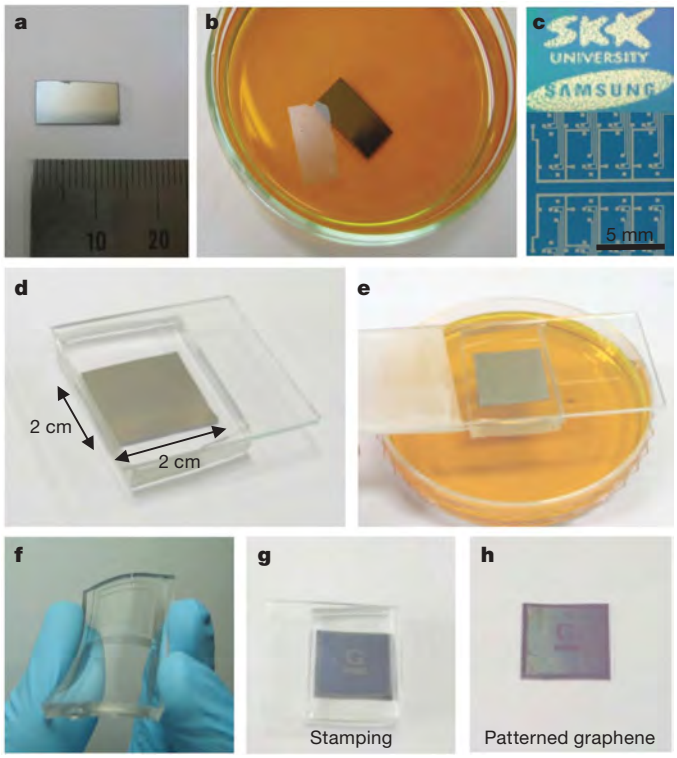


Figure 3 | Transfer processes for large-scale graphene films. **a**, A centimetre-scale graphene film grown on a Ni(300 nm)/SiO_2(300 nm)/Si substrate. **b**, A floating graphene film after etching the nickel layers in 1 M $FeCl_3$ aqueous solution. After the removal of the nickel layers, the floating graphene film can be transferred by direct contact with substrates. **c**, Various shapes of graphene films can be synthesized on top of patterned nickel layers. **d**, **e**, The dry-transfer method based on a PDMS stamp is useful in transferring the patterned graphene films. After attaching the PDMS substrate to the graphene (**d**), the underlying nickel layer is etched and removed using $FeCl_3$ solution (**e**). **f**, Graphene films on the PDMS substrates are transparent and flexible. **g**, **h**, The PDMS stamp makes conformal contact with a silicon dioxide substrate. Peeling back the stamp (**g**) leaves the film on a SiO_2 substrate (**h**).

Etching nickel substrate layers and transferring isolated graphene films to other substrates is important for device applications. Usually, nickel can be etched by strong acid such as HNO_3, which often produces hydrogen bubbles and damages the graphene. In our work, an aqueous iron (III) chloride ($FeCl_3$) solution (1 M) was used as an oxidizing etchant to remove the nickel layers. The net ionic equation of the etching reaction can be represented as follows:

$$2Fe^{3+}(aq) + Ni(s) \rightarrow 2Fe^{2+}(aq) + Ni^{2+}(aq)$$

This redox process slowly etches the nickel layers effectively within a mild pH range without forming gaseous products or precipitates. In a few minutes, the graphene film separated from the substrate floats on the surface of the solution (Fig. 3a, b), and the film is then ready to be transferred to any kind of substrate. Use of buffered oxide etchant (BOE) or hydrogen fluoride solution removes silicon dioxide layers, so the patterned graphene and the nickel layer float together on the solution surface. After transfer to a substrate, further reaction with BOE or hydrogen fluoride solution completely removes the remaining nickel layers (Supplementary Fig. 5).

We also develop a dry-transfer process for the graphene film using a soft substrate such as polydimethylsiloxane (PDMS) stamp[24]. Here we first attach the PDMS stamp to the CVD-grown graphene film on the nickel substrate (Fig. 3d). The nickel substrate can be etched away using $FeCl_3$ as described above, leaving the adhered graphene film on the PDMS substrate (Fig. 3e). By using the pre-patterned nickel substrate (Fig. 3c), we can transfer various sizes and shapes of graphene film to an arbitrary substrate. This dry-transfer process turns out to be very useful in making large-scale graphene electrodes and devices without additional lithography processes (Fig. 3f–h). Microscopically, these few-layer transferred graphene films often show linear crack patterns with an angle of 60° or 120°, indicating a particular crystallographic edge with large crystalline domains (Supplementary Fig. 1b)[25]. In addition, the Raman spectra measured for graphene films on nickel substrates show a strongly suppressed defect-related D-band peak (Supplementary Fig. 3). This D peak grows only slightly after the transfer process (Fig. 2e), indicating overall good quality of the resulting graphene film. Further optimization of the transfer process with substrate control makes possible transfer yields approaching 99% (Supplementary Table 1).

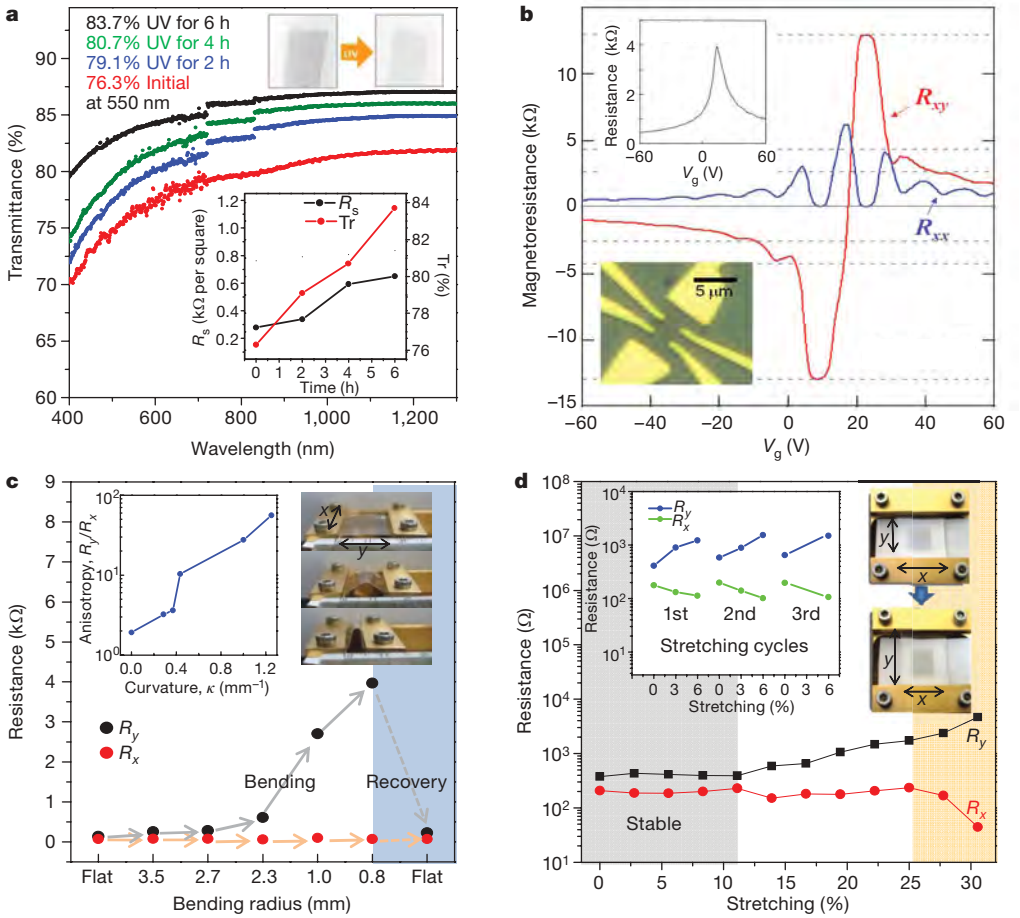


Figure 4 | Optical and electrical properties of the graphene films. **a**, Transmittance of the graphene films on a quartz plate. The discontinuities in the absorption curves arise from the different sensitivities of the switching detectors. The upper inset shows the ultraviolet (UV)-induced thinning and the consequent enhancement of transparency. The lower inset shows the changes in transmittance, Tr, and sheet resistance, R_s, as functions of ultraviolet illumination time. **b**, Electrical properties of monolayer graphene devices showing the half-integer quantum Hall effect and high electron mobility. The upper inset shows a four-probe electrical resistance measurement on a monolayer graphene Hall bar device (lower inset) at 1.6 K. We apply a gate voltage, V_g, to the silicon substrate to control the charge density in the graphene sample. The main panel shows longitudinal (R_{xx}) and transverse (R_{xy}) magnetoresistances measured in this device for a magnetic field $B = 8.8$ T. The monolayer graphene quantum Hall effect is clearly observed, showing the plateaux with filling factor $\nu = 2$ at $R_{xy} = (2e^2/h)^{-1}$ and zeros in R_{xx}. (Here e is the elementary charge and h is Planck's constant.) Quantum Hall plateaux (horizontal dashed lines) are developing for higher filling factors. **c**, Variation in resistance of a graphene film transferred to a ~0.3-mm-thick PDMS/PET substrate for different distances between holding stages (that is, for different bending radii). The left inset shows the anisotropy in four-probe resistance, measured as the ratio, R_y/R_x, of the resistances parallel and perpendicular to the bending direction, y. The right inset shows the bending process. **d**, Resistance of a graphene film transferred to a PDMS substrate isotropically stretched by ~12%. The left inset shows the case in which the graphene film is transferred to an unstretched PDMS substrate. The right inset shows the movement of holding stages and the consequent change in shape of the graphene film.

For the macroscopic transport electrode application, the optical and electrical properties of $1 \times 1\ cm^2$ graphene films were respectively measured by ultraviolet–visible spectrometer and four-probe Van der Pauw methods (Fig. 4a, b). We measured the transmittance using an ultraviolet–visible spectrometer (UV-3600, Shimazdu) after transferring the floating graphene film to a quartz plate (Fig. 4a). In the visible range, the transmittance of the film grown on a 300-nm-thick nickel layer for 7 min is ~80%, a value similar to those found for previously studied assembled films[2,3]. Because the transmittance of an individual graphene layer is ~2.3% (ref. 26), this transmittance value indicates that the average number of graphene layers is six to ten. The transmittance can be increased to ~93% by further reducing the growth time and nickel thickness, resulting in a thinner graphene film (Supplementary Fig. 1). Ultraviolet/ozone etching (ultraviolet/ozone cleaner, 60 W, BioForce) is also useful in controlling the transmittance in an ambient condition (Fig. 4a, upper inset). Indium electrodes were deposited on each corner of the square (Fig. 4a, lower inset) to minimize contact resistance. The minimum sheet resistance is ~280 Ω per square, which is ~30 times smaller than the lowest sheet resistance measured on assembled films[2,3]. The values of sheet resistance increase with the ultraviolet/ozone treatment time, in accordance with the decreasing number of graphene layers (Fig. 4a).

For microelectronic application, the mobility of the graphene film is critical. To measure the intrinsic mobility of a single-domain graphene sample, we transferred the graphene samples from a PDMS stamp to a degenerate doped silicon wafer with a 300-nm-deep thermally grown oxide layer. Monolayer graphene samples were readily located on the substrate from the optical contrast[26] and their identification was subsequently confirmed by Raman spectroscopy[22]. Electron-beam lithography was used to make multi-terminal devices (Fig. 4b, lower inset). Notably, the multi-terminal electrical measurements showed that the electron mobility is ~3,750 $cm^2\,V^{-1}\,s^{-1}$ at a carrier density of ~$5 \times 10^{12}\ cm^{-2}$ (Fig. 4b). For a high magnetic field of 8.8 T, we observe the half-integer quantum Hall effect (Fig. 4b) corresponding to monolayer graphene[4,5], indicating that the quality of CVD-grown graphene is comparable to that of mechanically cleaved graphene (Supplementary Fig. 6)[6].

In addition to the good optical and electrical properties, the graphene film has excellent mechanical properties when used to make flexible and stretchable electrodes (Fig. 4c, d)[7]. We evaluated the foldability of the graphene films transferred to a polyethylene terephthalate (PET) substrate (thickness, ~100 μm) coated with a thin PDMS layer (thickness, ~200 μm; Fig. 4c) by measuring resistances with respect to bending radii. The resistances show little variation up to the bending radius of 2.3 mm (approximate tensile strain of 6.5%) and are perfectly recovered after unbending. Notably, the original resistance can be restored even for the bending radius of 0.8 mm (approximate tensile strain of 18.7%), exhibiting extreme mechanical stability in comparison with conventional materials used in flexible electronics[27].

The resistances of graphene films transferred to pre-strained and unstrained PDMS substrates were measured with respect to uniaxial tensile strain ranging from 0 to 30% (Fig. 4d). Similar to the results in the folding experiment, the transferred film on an unstrained substrate recovers its original resistance after stretching by ~6% (Fig. 4d, left inset). However, further stretching often results in mechanical failure. Thus, we tried to transfer the film to pre-strained substrates[28] to enhance the electromechanical stabilities by creating ripples similar to those observed in the growth process (Fig. 2c, inset; Supplementary Fig. 4). The graphene transferred to a longitudinally strained PDMS substrate does not show much enhancement, owing to the transverse strain induced by Poisson's effect[29]. To prevent this problem, the PDMS substrate was isotropically stretched by ~12% before transferring the film to it (Fig. 4d). Surprisingly, both longitudinal and transverse resistances (R_y and R_x) appear stable up to ~11% stretching and show only one order of magnitude change at ~25% stretching. We suppose that further uniaxial stretching can change the electronic band structures of graphene, leading to the modulation of the sheet resistance. These electromechanical properties thus show our graphene films to be not only the strongest[7] but also the most flexible and stretchable conducting transparent materials so far measured[26].

In conclusion, we have developed a simple method to grow and transfer high-quality stretchable graphene films on a large scale using CVD on nickel layers. The patterned films can easily be transferred to stretchable substrates by simple contact methods, and the number of graphene layers can be controlled by varying the thickness of the catalytic metals, the growth time and/or the ultraviolet treatment time. Because the dimensions of the graphene films are limited simply by the size of the CVD growth chamber, scaling up can be readily achieved, and the outstanding optical, electrical and mechanical properties of the graphene films enable numerous applications including use in large-scale flexible, stretchable, foldable transparent electronics[8,9,30].

Received 5 October; accepted 8 December 2008.
Published online 14 January 2009.

1. Geim, A. K. & Novoselov, K. S. The rise of graphene. *Nature Mater.* **6,** 183–191 (2007).
2. Li, X. *et al.* Highly conducting graphene sheets and Langmuir–Blodgett films. *Nature Nanotechnol.* **3,** 538–542 (2008).
3. Eda, G., Fanchini, G. & Chhowalla, M. Large-area ultrathin films of reduced graphene oxide as a transparent and flexible electronic material. *Nature Nanotechnol.* **3,** 270–274 (2008).
4. Novoselov, K. S. *et al.* Two-dimensional gas of massless Dirac fermions in graphene. *Nature* **438,** 197–200 (2005).
5. Zhang, Y., Tan, J. W., Stormer, H. L. & Kim, P. Experimental observation of the quantum Hall effect and Berry's phase in graphene. *Nature* **438,** 201–204 (2005).
6. Novoselov, K. S. *et al.* Electric field effect in atomically thin carbon films. *Science* **306,** 666–669 (2004).
7. Lee, C., Wei, X., Kysar, J. W. & Hone, J. Measurement of the elastic properties and intrinsic strength of monolayer graphene. *Science* **321,** 385–388 (2008).
8. Kim, D.-H. *et al.* Stretchable and foldable silicon integrated circuits. *Science* **320,** 507–511 (2008).
9. Sekitani, T. *et al.* A rubberlike stretchable active matrix using elastic conductors. *Science* **321,** 1468–1472 (2008).
10. Han, M. Y., Oezyilmaz, B., Zhang, Y. & Kim, P. Energy band gap engineering of graphene nanoribbons. *Phys. Rev. Lett.* **98,** 206805 (2007).
11. Bolotin, K. I. *et al.* Ultrahigh electron mobility in suspended graphene. *Solid State Commun.* **146,** 351–355 (2008).
12. Bunch, J. S. *et al.* Electromechanical resonators from graphene sheets. *Science* **315,** 490–493 (2008).
13. Ohta, T., Bostwick, A., Seyller, T., Horn, K. & Rotenberg, E. Controlling the electronic structure of bilayer graphene. *Science* **313,** 951–954 (2006).
14. Berger, C. *et al.* Electronic confinement and coherence in patterned epitaxial graphene. *Science* **312,** 1191–1196 (2006).
15. Sutter, P. W., Flege, J.-I. & Sutter, E. A. Epitaxial graphene on ruthenium. *Nature Mater.* **7,** 406–411 (2008).
16. Dikin, D. A. *et al.* Preparation and characterization of graphene oxide paper. *Nature* **448,** 457–460 (2007).
17. Stankovich, S, *et al.* Graphene-based composite materials. *Nature* **442,** 282–286 (2006).
18. Li, D., Muller, M. B., Gilje, S., Kaner, R. B. & Wallace, G. G. Processable aqueous dispersions of graphene nanosheets. *Nature Nanotechnol.* **3,** 101–105 (2008).
19. Obraztsov, A. N., Obraztsova, E. A., Tyurnina, A. V. & Zolotukhin, A. A. Chemical vapor deposition of thin graphite films of nanometer thickness. *Carbon* **45,** 2017–2021 (2007).
20. Yu, Q, *et al.* Graphene segregated on Ni surfaces and transferred to insulators. *Appl. Phys. Lett.* **93,** 113103 (2008).
21. Reina, A. *et al.* Large area, few-layer graphene films on arbitrary substrates by chemical vapor deposition. *Nano Lett.* article ASAP at ⟨http://pubs.acs.org/doi/abs/10.1021/nl801827v⟩ (2008).
22. Ferrari, A. C. *et al.* Raman spectrum of graphene and graphene layers. *Phys. Rev. Lett.* **97,** 187401 (2006).
23. Khang, D.-Y. *et al.* Individual aligned single-wall carbon nanotubes on elastomeric substrates. *Nano Lett.* **8,** 124–130 (2008).
24. Yang, P. *et al.* Mirrorless lasing from mesostructured waveguides patterned by soft lithography. *Science* **287,** 465–467 (2000).
25. Li, X., Wang, X., Zhang, L., Lee, S. & Dai, H. Chemically derived, ultrasmooth graphene nanoribbon semiconductors. *Science* **319,** 1229–1232 (2008).
26. Nair, R. R. *et al.* Fine structure constant defines visual transparency of graphene. *Science* **320,** 1308 (2008).
27. Lewis, J. Material challenge for flexible organic devices. *Mater. Today* **9,** 38–45 (2006).
28. Sun, Y., Choi, W. M., Jiang, H., Huang, Y. Y. & Rogers, J. A. Controlled buckling of semiconductor nanoribbons for stretchable electronics. *Nature Nanotechnol.* **1,** 201–207 (2006).

29. Khang, D.-Y., Jiang, H., Huang, Y. & Rogers, J. A. A stretchable form of single-crystal silicon for high-performance electronics on rubber substrates. *Science* **311,** 208–212 (2006).
30. Ko, H. C. *et al.* A hemispherical electronic eye camera based on compressible silicon optoelectronics. *Nature* **454,** 748–753 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. H. Han, J. H. Kim, H. Lim, S. K. Bae and H.-J. Shin for assisting in graphene synthesis and analysis. This work was supported by the Korea Science and Engineering Foundation grant funded by the Korea Ministry for Education, Science and Technology (Center for Nanotubes and Nanostructured Composites R11-2001-091-00000-0), the Global Research Lab programme (Korea Foundation for International Cooperation of Science and Technology), the Brain Korea 21 project (Korea Research Foundation) and the information technology research and development programme of the Korea Ministry of Knowledge Economy (2008-F024-01).

Author Contributions B.H.H. planned and supervised the project; J.-Y.C. supported and assisted in supervision on the project; S.Y.L, J.M.K. and K.S.K. advised on the project; K.S.K. and B.H.H. designed and performed the experiments; B.H.H., P.K., J.-H.A and K.S.K. analysed data and wrote the manuscript; Y.Z. and P.K. made the quantum Hall devices and the measurements; and H.J. and J.-H.A. helped with the transfer process and the electromechanical analyses.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to B.H.H. (byunghee@skku.edu) or J.-Y.C. (jaeyoung88.choi@samsung.com).

LETTERS

Holocene oscillations in temperature and salinity of the surface subpolar North Atlantic

David J. R. Thornalley[1]†, Harry Elderfield[1] & I. Nick McCave[1]

The Atlantic meridional overturning circulation (AMOC) transports warm salty surface waters to high latitudes, where they cool, sink and return southwards at depth. Through its attendant meridional heat transport, the AMOC helps maintain a warm northwestern European climate, and acts as a control on the global climate. Past climate fluctuations during the Holocene epoch (~11,700 years ago to the present) have been linked with changes in North Atlantic Ocean circulation[1,2]. The behaviour of the surface flowing salty water that helped drive overturning during past climatic changes is, however, not well known. Here we investigate the temperature and salinity changes of a substantial surface inflow to a region of deep-water formation throughout the Holocene. We find that the inflow has undergone millennial-scale variations in temperature and salinity (~3.5 °C and ~1.5 practical salinity units, respectively) most probably controlled by subpolar gyre dynamics. The temperature and salinity variations correlate with previously reported periods of rapid climate change[3]. The inflow becomes more saline during enhanced freshwater flux to the subpolar North Atlantic. Model studies predict a weakening of AMOC in response to enhanced Arctic freshwater fluxes[4], although the inflow can compensate on decadal timescales by becoming more saline[5]. Our data suggest that such a negative feedback mechanism may have operated during past intervals of climate change.

The AMOC is a critical component of the Earth's climate system, redistributing heat and partitioning carbon between the surface and deep ocean reservoirs. The surface limb of the AMOC consists of the warm, saline, surface North Atlantic Current (NAC) that flows northeastwards across the North Atlantic into the Nordic seas (hereafter referred to as the Atlantic inflow), passing between the subpolar and subtropical gyres, from which it draws water[6] (Fig. 1). On entering the Nordic seas, cooling promotes the formation of deep water, which overflows the Greenland–Scotland ridge and returns southwards as a major component of North Atlantic Deep Water[6]—the deep-water limb of the AMOC. Here we investigate the hydrography of the Atlantic inflow throughout the Holocene, a period that before 8 kyr ago includes enhanced freshwater fluxes due to ice-sheet disintegration.

The Holocene has experienced considerable climatic variability on decadal[6,7] to millennial[1,2,8,9] timescales, notwithstanding the almost constant isotopic record in the Greenland ice cores[10]. Instrumental, historical and proxy data have documented rapid and large changes in the position of the subpolar front (intra-annual fluctuations of up to 300 km and millennial excursions of up to 500 km), bringing cold, fresh, ice-bearing waters to the coasts of Iceland[6,7,9] These climatic fluctuations are linked with important cultural and socio-economic changes throughout northwestern Europe[9] and, through teleconnections, globally[3].

We reconstruct the temperature and salinity of the Atlantic inflow using paired Mg/Ca–$\delta^{18}O$ (see Methods Summary) measurements on two species of planktonic foraminifera, *Globigerina bulloides* and *Globorotalia inflata*, following the procedures of ref. 11. The oxygen-isotopic composition of planktonic foraminiferal calcite depends upon both calcification temperature and the ambient seawater $\delta^{18}O$ ($\delta^{18}O_{sw}$). The Mg/Ca ratio of planktonic foraminiferal calcite is controlled primarily by calcification temperature[12]. Combined Mg/Ca–$\delta^{18}O$ measurements therefore allow the reconstruction of temperature and $\delta^{18}O_{sw}$; palaeosalinity can then be estimated using modern $\delta^{18}O_{sw}$–salinity relationships, although it is uncertain how the $\delta^{18}O_{sw}$–salinity relationship changes through time (see Methods Summary for errors)[13]. By examining species with different depth habitats, reconstructions can sample both the surface and volumetrically significant waters below. The stable isotope compositions of both species reflect conditions during late spring and early summer[14]. *G. bulloides* occupies the seasonal mixed layer, typically 0–50 m below the surface (ref. 14). *G. inflata* calcifies at the base of the seasonal thermocline[15], (~100–200 m)[14], in waters cooled during winter convection. Shallow temperature and salinity gradients below the

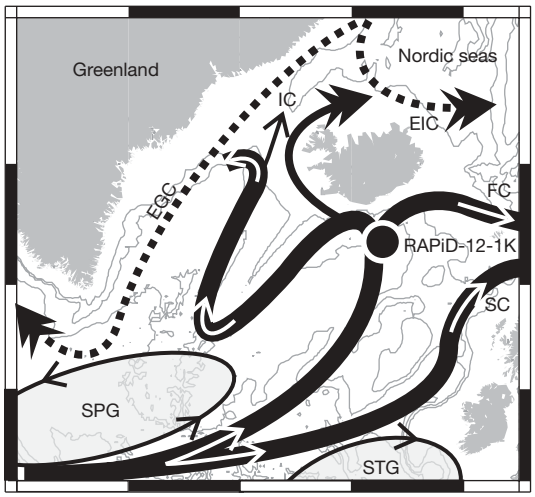


Figure 1 | Map of study area showing the main features of the surface circulation in the northeast North Atlantic[16]. Location of core RAPiD-12-1K (62° 05.43′ N, 17° 49.18′ W; 1,938-m water depth) is marked with a black circle. Continuous arrows show the main branches of the North Atlantic Current, namely the Irminger Current (IC, 1 Sv), the Faroe Current (FC, 3.3 Sv) and the Shetland Current (SC, 3.7 Sv)[6], which draw water from the subpolar gyre (SPG) and the subtropical gyre (STG). Dashed lines show the East Greenland Current (EGC, at least 1.3 Sv across the ridge[6]) and the East Icelandic Current (EIC).

[1]The Godwin Laboratory for Palaeoclimate Research, Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ, UK. †Present address: School of Earth and Ocean Sciences, Cardiff University, Main Building, Park Place, Cardiff CF10 3YE, UK.

thermocline may reduce the impact of habitat depth migration on reconstructions based on *G. inflata*.

Records were made using sediment core RAPiD-12-1K (62° 05.43′ N, 17° 49.18′ W; 1,938-m water depth) from the South Iceland rise (Fig. 1). Sedimentation rates average 115 cm kyr^{-1} from 12 to 8 kyr ago and 23 cm kyr^{-1} from 8 kyr ago to the present (dated by ^{14}C accelerator mass spectrometry (Supplementary Methods)). The core lies under the path of the NAC where it bifurcates to form the Irminger and Faroe currents[16], although instrumental and historical records also document episodes of cold ice-bearing subpolar waters from the north reaching the site[7,9]. Modern hydrographic measurements indicate a well-mixed upper water column down to at least 600 m, with a nearly constant salinity of 35.2–35.3 practical salinity units (p.s.u.) and a temperature of ~8 °C, and seasonal warming of the upper 50–100 m to 11.5 °C (ref. 17).

The Mg/Ca and δ^{18}O data for *G. bulloides* reveal millennial salinity variations of ~0.5 p.s.u. superimposed upon a trend of increasing near-surface water salinity from ~9 kyr ago to the present (Fig. 2). Temperatures remain nearly constant at 10–11 °C, reflecting a similar seasonal warming of near-surface waters. The early Holocene between 11 and 8 kyr ago is characterized by low salinities, fluctuating around ~34 p.s.u. The longer timescale trends in near-surface salinity are consistent with nearby Mg/Ca-based near-surface salinity reconstructions[8], which may be caused by some or all of the following: net Atlantic Ocean salinity changes related to the gradual migration of the intertropical convergence zone[18]; early-Holocene input of light-$\delta^{18}O_{sw}$ deglacial melt water to the North Atlantic[8]; and changes in freshwater export from the Arctic Ocean. Centennial to millennial freshening of the near-surface water most likely reflects the southward advance of the subpolar front, as in the 1960s when atmospheric changes (North Atlantic Oscillation minimum conditions) resulted in more northerly winds exporting sea ice southwards from the Nordic seas[7]. In response to surface freshening, it is possible that *G. bulloides* migrated to a deeper, more saline, environment and that the freshening is underestimated.

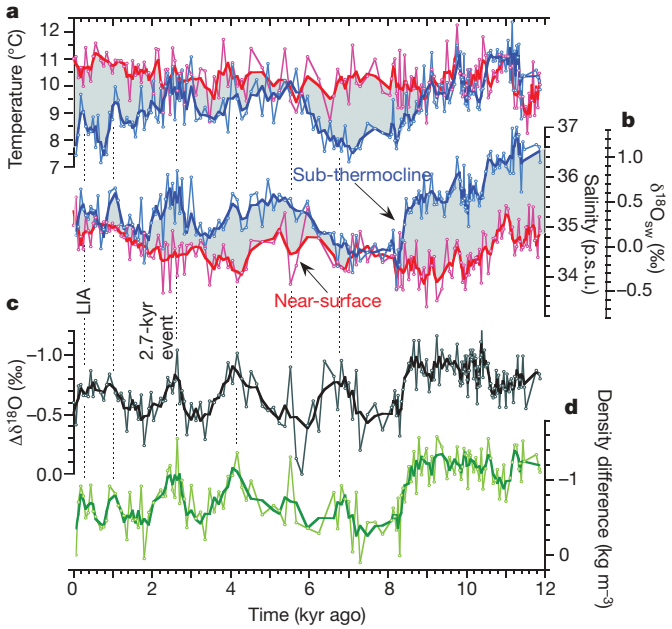


Figure 2 | Proxy records for RAPiD-12-1K. a, **b**, Mg/Ca-based temperatures (**a**) and salinity estimates derived from paired Mg/Ca–δ^{18}O measurements (**b**), for near-surface (*G. bulloides*, red) and sub-thermocline (*G. inflata*, blue) waters. Also shown is a scale bar for $\delta^{18}O_{sw}$ values, corrected for whole-ocean ice-volume changes. **c**, **d**, Proxies for upper-water-column stratification (stratification increases upwards) based on the δ^{18}O difference between *G. bulloides* and *G. inflata* (**c**) and the inferred water density difference between *G. bulloides* and *G. inflata* (**d**), calculated using derived temperatures and salinities. Three-point running means shown in bold. LIA, Little Ice Age.

During the early Holocene, the sub-thermocline (*G. inflata*) data show that temperatures were similar to that of the fresh near-surface layer but that salinity was greater. Between 9 and 8 kyr ago, strong sub-thermocline cooling and freshening occurred, with the result that, between 8 and 7 kyr ago, the structure of the upper water column was similar, but fresher, than it is now. During this transition, the glacial freshwater discharge event of 8.2 kyr ago can be recognized as a 0.5 p.s.u. sub-thermocline freshening, similar in amplitude to salinity variations reported in previous studies[8]. Sub-thermocline temperature and salinity oscillate throughout the remainder of the Holocene. Warm saline sub-thermocline conditions are centred at 0.3, 1.0, 2.7 and 5.0 kyr ago, coinciding with known climatic perturbations in the North Atlantic region, for example the Little Ice Age and the cold event 2.7 kyr ago (Fig. 2).

Modern controls on regional salinity identified in ref. 5 include (1) local air–sea fluxes of freshwater, (2) variations in salinity of the subpolar gyre (SPG) or (3) the subtropical gyre (STG), and (4) dynamic changes in the relative contributions from the two gyres. Mechanism (1) does not explain the sub-thermocline changes, because air–sea fluxes typical for this region[6] would cause a much steeper temperature–salinity gradient than measured (Supplementary Discussion). Furthermore, the upper water column properties are set at source and during advection across the North Atlantic basin, so the magnitude of the salinity variations observed would require extreme changes in air–sea fluxes in the Caribbean and over the entire northern North Atlantic. Mechanism (2) is not responsible, because salinity estimates from the Labrador Sea[19], located within the SPG, show fresh conditions in contrast to the saline conditions south of Iceland. Mechanism (3) is of potential importance considering the amplitude of subsurface STG salinity variability over the past 110 years[20]. However, there is no significant correlation between our south Iceland record and STG salinity records, namely late-Holocene Florida Current surface salinity records[21], and Holocene Mg/Ca–δ^{18}O-based salinity estimates for STG mode waters, derived from the deep-dwelling foraminifer *Globorotalia truncatulinoides*[22]. Further confirmation that mechanism (3) is not a significant factor will require additional subsurface STG salinity records to be produced.

Consistent with modern mechanisms[5], we conclude that throughout the Holocene, the salinity of water below the near-surface layer south of Iceland is primarily controlled by mechanism (4), that is, by the proportion of water being drawn from either the cold, fresh SPG or the warm, saline STG, which has been shown to depend strongly on the dynamics of the SPG[4]. Strong SPG circulation strength results in a more East–West-oriented SPG, which therefore contributes more water to the Atlantic inflow, making it fresher. Conversely, weak SPG circulation strength results in a more North–South-oriented SPG, which contributes less water to the Atlantic inflow, making it saltier. The strength of the SPG circulation can depend on the local wind stress and/or the baroclinic circulation driven by buoyancy forcing (associated with deep convection)[23]. Freshwater input to the Labrador Sea prevents deep convection[24], thereby reducing SPG circulation and the SPG influence south of Iceland.

Figure 3 illustrates that existing records are consistent with the SPG strength mechanism: fresh surface conditions in the Labrador Sea[19] coincide with a reduction in influence of the SPG in the eastern subtropical North Atlantic[25] and more saline conditions south of Iceland. These changes occur during recognized periods of global rapid climate change, involving ocean and atmosphere reorganizations[3]. Early-Holocene freshening of the Labrador Sea was most likely driven by deglacial meltwater input, and/or enhanced freshwater flux, through the East Greenland Current, during more meridional atmospheric circulation[26]. Late-Holocene saline intervals south of Iceland, which indicate weak SPG circulation, are not accompanied by changes in Labrador Sea salinity. This suggests that weakened SPG circulation may have been caused by decreased wind stress, rather

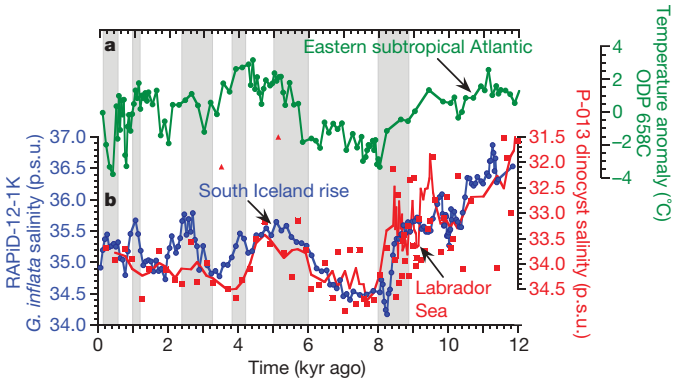


Figure 3 | Records of changing gyre properties. a, Linear detrended temperature anomalies obtained from ocean sediment core ODP 658C, west coast of Africa (green line and points)[25]. Warmer temperatures are caused by either decreased upwelling or decreased advection of subpolar waters into the eastern Atlantic during weak subpolar gyre circulation. **b**, Salinity estimates from *G. inflata* for RAPiD-12-1K (blue line and points), and dinocyst-assemblage salinity estimates from ocean sediment core P-013, central Labrador Sea (subpolar gyre)[19] on a reversed axis (red squares; red triangles are outliers not included in the three-point mean (red line)). Grey shaded regions are periods of global rapid climate change[3].

than by enhanced freshwater flux. Late-Holocene freshening in the Labrador Sea was possibly also less pervasive and of shorter duration, and thus not recorded by dinocyst assemblages[19].

Critically, all periods of enhanced surface freshening in the Labrador Sea are accompanied by more saline conditions south of Iceland. SPG dynamics can therefore act as a negative feedback, stabilizing the AMOC to freshwater input. The potential importance of SPG dynamics on the AMOC can be illustrated by examining a recent modelling study of Holocene AMOC variability[27]. In this study, millennial oscillations of the AMOC are caused by convective shutdown in the Labrador Sea, and its upstream surface-water linkage to the Nordic seas[27]. Palaeo-oceanographic reconstructions show that convective shutdown has occurred on several occasions throughout the Holocene[24]. This will have reduced SPG circulation strength, producing a more saline Atlantic inflow to the Nordic seas, which eventually fed through to the Labrador Sea (via the East Greenland Current) and restarted convection. These oscillations may have been controlled by a weak external driver such as solar variability[27]. It has been suggested that weakening of the AMOC occurred 6–5 kyr ago and 2.8 kyr ago, during southward advance of sea ice and a change in atmospheric circulation[28]. The increased salinity of the Atlantic inflow observed during these periods may have limited the reduction, or helped restart stronger AMOC.

Holocene variability in Atlantic inflow properties, and the southward migration of the subpolar front, cause fluctuating density stratification of the upper water column south of Iceland (Fig. 2). The records show a stratified upper ocean during the early Holocene with an abrupt switch to well-mixed waters ~8.4 kyr ago, followed by quasi-periodic stratification events every ~1,500 yr. This suggests that surface circulation was fixed in one mode of operation before ~8.4 kyr ago, perhaps owing to the deglacial input of melt water to the SPG. Later, with reduced freshwater input, the system oscillated between two modes of operation, involving strong and weak SPG circulation. This threshold behaviour is similar to that displayed by the model of ref. 27. Spectral analysis of the density stratification record from 8.4 kyr ago to present, using confidence limits of 90%, shows one broad peak centred at 1,500 yr (Supplementary Notes), consistent with other North Atlantic studies[1,2]. This cyclicity has been attributed to ocean dynamics[29] and the data here confirm that oceanic factors underlie the oscillations.

The importance of the salinity balance in the North Atlantic is well established, with the transfer of subtropical salinity to high latitudes invoked to precondition and help restart deep overturning during rapid climate fluctuations[30]. The Atlantic inflow is of paramount importance in the transport of salt from low to high latitudes. We have shown that this transport has undergone large-amplitude millennial variability modulated by SPG dynamics. Although further confirmation of this mechanism will require additional subsurface North Atlantic salinity records, this critical process should be included when examining the dynamics of the AMOC and its involvement in climate changes.

METHODS SUMMARY

Around 20–30 tests of *G. bulloides* and *G. inflata* (300–355-μm fraction) were analysed for $\delta^{18}O$ and Mg/Ca ratios following published methods[11], screening for contaminating ferromanganese overgrowths, clay minerals and silicate particles. Analytical precision of Mg/Ca ratios based on replicates of foraminiferal standards is 3%. Shell weights show Mg/Ca ratios are not affected by dissolution.

Oxygen isotope ratios ($\delta^{18}O = (^{18}O/^{16}O)_{sample}/(^{18}O/^{16}O)_{standard} - 1$) were determined by means of gas-source mass spectrometry relative to the Vienna PeeDee belemnite standard. Analytical precision based on long-term replicates is better than 0.08‰.

An exponent of $A = 0.10$ was used[12] in the equation for Mg/Ca: $Mg/Ca = B\exp(A \times T)$. Core-top Mg/Ca values were calibrated to modern hydrographic data, yielding values for B of 0.794 and 0.675 for *G. bulloides* and *G. inflata*, respectively. T denotes temperature.

The estimated standard deviations in absolute T, $\delta^{18}O_{sw}$ and salinity, S, are 1.3 °C, 0.32‰ and 0.8 p.s.u., respectively[13]. The estimated standard deviations in relative T, $\delta^{18}O_{sw}$ and S are 1.0 °C, 0.26‰ and 0.46 p.s.u., respectively. These estimates include measurement errors, sample heterogeneity, carbonate ion effects and ice-volume effect uncertainty but ignore calibration errors, which should be more constant down-core, and changes in the S–$\delta^{18}O_{sw}$ relationship, which on a regional and larger scale should affect both species similarly. The introduction of glacial melt water will result in light $\delta^{18}O_{sw}$ values and anomalously low salinities may be reconstructed. The import and subsequent melting of sea ice south of Iceland will freshen the water column with only a minor change in $\delta^{18}O_{sw}$; $\delta^{18}O_{sw}$ reconstructions may therefore underestimate the surface freshening. Comparison between the *G. bulloides* and *G. inflata* data help constrain the relative changes through time.

Received 17 March; accepted 9 December 2008.

1. Bond, G. *et al.* Persistent solar influence on north Atlantic climate during the Holocene. *Science* **294,** 2130–2136 (2001).
2. Bianchi, G. G. & McCave, I. N. Holocene periodicity in North Atlantic climate and deep-ocean flow south of Iceland. *Nature* **397,** 515–517 (1999).
3. Mayewski, P. A. *et al.* Holocene climate variability. *Quat. Res.* **62,** 243–255 (2004).
4. Cubasch, U. & Meehl, G. in *Climate Change 2001: The Scientific Basis* (eds Houghton, J. T. *et al.*) 525–582 (Cambridge Univ. Press, 2001).
5. Hátún, H., Sando, A. B., Drange, H., Hansen, B. & Valdimarsson, H. Influence of the Atlantic subpolar gyre on the thermohaline circulation. *Science* **309,** 1841–1844 (2005).
6. Hansen, B. & Østerhus, S. North Atlantic-Nordic Seas exchanges. *Prog. Oceanogr.* **45,** 109–208 (2000).
7. Blindheim, J. & Østerhus, S. The Nordic Seas, main oceanographic features. *Geophys. Monogr.* **158,** 11–37 (2005).
8. Came, R. E., Oppo, D. W. & McManus, J. F. Amplitude and timing of temperature and salinity variability in the subpolar North Atlantic over the past 10 ky. *Geology* **35,** 315–318 (2007).
9. Lamb, H. H. Climatic variation and changes in the winds and ocean circulation: The Little Ice Age and the Northeast Atlantic. *Quat. Res.* **11,** 1–20 (1979).
10. Andersen, K. K. *et al.* High-resolution record of Northern Hemisphere climate extending into the last interglacial period. *Nature* **431,** 147–151 (2004).
11. Barker, S., Greaves, M. & Elderfield, H. A study of cleaning procedures used for foraminiferal Mg/Ca paleothermometry. *Geochem. Geophys. Geosyst.* **4,** 8407–8427 (2003).
12. Barker, S., Cacho, I., Benway, H. & Tachikawa, K. Planktonic foraminiferal Mg/Ca as a proxy for past oceanic temperatures: a methodological overview and data compilation for the Last Glacial Maximum. *Quat. Sci. Rev.* **24,** 821–834 (2005).
13. Schmidt, G. A. Error analysis of paleosalinity calculations. *Paleoceanography* **14,** 422–429 (1999).
14. Ganssen, G. M. & Kroon, D. The isotopic signature of planktonic foraminifera from NE Atlantic surface sediments: implications for the reconstruction of past oceanic conditions. *J. Geol. Soc. Lond.* **157,** 693–699 (2000).
15. Cléroux, C., Cortijo, E., Duplessy, J.-C. & Zahn, R. Deep-dwelling foraminifera as thermocline temperature recorders. *Geochem. Geophys. Geosyst.* **8,** doi:10.1029/2006GC001474 (2007).

16. Orvik, K. A. & Niiler, P. Major pathways of Atlantic water in the northern North Atlantic and Nordic Seas toward Arctic. *Geophys. Res. Lett.* **29,** 1896–1900 (2002).
17. US National Oceanic and Atmospheric Administration. NODC (Levitus) World Ocean Atlas 1998. *Earth System Research Laboratory, Physical Sciences Division* ⟨http://www.cdc.noaa.gov/⟩ (1998).
18. Haug, G. H., Hughen, K. A., Sigman, D. M., Peterson, L. & Röhl, U. Southward migration of the intertropical convergence zone through the Holocene. *Science* **293,** 1304–1308 (2001).
19. Solignac, S., de Vernal, A. & Hillaire-Marcel, C. Holocene sea-surface conditions in the North Atlantic - contrasted trends and regimes in the western and eastern sectors (Labrador Sea vs. Iceland Basin). *Quat. Sci. Rev.* **23,** 319–334 (2004).
20. Rosenheim, B. E., Swart, P. K., Thorrold, S. R., Eisenhauer, A. & Willenz, P. Salinity change in the subtropical Atlantic: Secular increase and teleconnections to the North Atlantic Oscillation. *Geophys. Res. Lett.* **29,** doi:10.1029/2004GL021499 (2005).
21. Lund, D. C. & Curry, W. B. Late Holocene variability in Florida current surface density: Patterns and possible causes. *Paleoceanography* **19,** doi:10.1029/2004PA001008 (2004).
22. Cléroux, C. *et al.* Upper water column hydrology changes off Cape Hatteras and Gulf Stream activity over the Holocene. *Eos* **88** (Fall meeting), abstr. PP13B–1276.
23. Häkkinen, S. & Rhines, P. B. Decline of subpolar North Atlantic circulation during the 1990s. *Science* **304,** 555–559 (2004).
24. Hillaire-Marcel, C., de Vernal, A., Bilodeau, G. & Weaver, A. Absence of deep-water formation in the Labrador Sea during the last interglacial period. *Nature* **410,** 1073–1077 (2001).
25. deMenocal, P. B., Ortiz, J., Guilderson, T. & Sarnthein, M. Coherent high- and low-latitude climate variability during the Holocene Warm Period. *Science* **288,** 2198–2202 (2000).
26. Dickson, R., Lazier, J., Meincke, J., Rhines, P. & Swift, J. Long-term coordinated changes in the convective activity of the North Atlantic. *Prog. Oceanogr.* **38,** 241–295 (1996).
27. Schulz, M., Prange, M. & Klocker, A. Low-frequency oscillations of the Atlantic Ocean meridional overturning circulation in a coupled climate model. *Clim. Past* **3,** 97–107 (2007).
28. Oppo, D. W., McManus, J. F. & Cullen, J. L. Deepwater variability in the Holocene epoch. *Nature* **422,** 277–278 (2003).
29. Debret, M. *et al.* The origin of the 1500-year climate cycles in Holocene North-Atlantic records. *Clim. Past* **3,** 569–575 (2007).
30. Schmidt, M. W., Vautravers, M. J. & Spero, H. J. Rapid subtropical North Atlantic salinity oscillations across Dansgaard-Oeschger cycles. *Nature* **443,** 561–564 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the crew of RV *Charles Darwin 159*; M. Greaves, A. Huckle and L. Booth for laboratory assistance; J. Rolfe and M. Hall for stable isotope analyses; J. Hillier for the Atlantic base map; and S. Crowhurst, T. Dokken, M. Schulz and L. Skinner for discussions. Radiocarbon dates were run by the UK Natural Environment Research Council (NERC) radiocarbon laboratory. Labrador Sea data was provided by A. de Vernal. Funding was provided by the NERC Rapid Climate Change programme.

Author Contributions H.E. and I.N.M. were responsible for initiating the study, and D.J.R.T. collected data, performed analyses and interpreted data. The manuscript was written by D.J.R.T. H.E. and I.N.M. contributed equally to the study. All authors contributed to the work at sea on RV *Charles Darwin 159*, discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to D.J.R.T. (d.thornalley@cantab.net).

LETTERS

Giant boid snake from the Palaeocene neotropics reveals hotter past equatorial temperatures

Jason J. Head[1], Jonathan I. Bloch[2], Alexander K. Hastings[2], Jason R. Bourque[2], Edwin A. Cadena[2,3], Fabiany A. Herrera[2,3], P. David Polly[4] & Carlos A. Jaramillo[3]

The largest extant snakes live in the tropics of South America and southeast Asia[1–3] where high temperatures facilitate the evolution of large body sizes among air-breathing animals whose body temperatures are dependant on ambient environmental temperatures (poikilothermy)[4,5]. Very little is known about ancient tropical terrestrial ecosystems, limiting our understanding of the evolution of giant snakes and their relationship to climate in the past. Here we describe a boid snake from the oldest known neotropical rainforest fauna from the Cerrejón Formation (58–60 Myr ago) in northeastern Colombia. We estimate a body length of 13 m and a mass of 1,135 kg, making it the largest known snake[6–9]. The maximum size of poikilothermic animals at a given temperature is limited by metabolic rate[4], and a snake of this size would require a minimum mean annual temperature of 30–34 °C to survive. This estimate is consistent with hypotheses of hot Palaeocene neotropics with high concentrations of atmospheric CO_2 based on climate models[10]. Comparison of palaeotemperature estimates from the equator to those from South American mid-latitudes indicates a relatively steep temperature gradient during the early Palaeogene greenhouse, similar to that of today. Depositional environments and faunal composition of the Cerrejón Formation indicate an anaconda-like ecology for the giant snake, and an earliest Cenozoic origin of neotropical vertebrate faunas.

Serpentes Linnaeus 1758
Boidae Gray 1825
Boinae Gray 1825
Titanoboa cerrejonensis gen. et sp. nov.

Etymology. The generic name combines '*Titan*' (Greek, giant) with '*Boa*', type genus for Boinae. The specific name refers to the Cerrejón region, Guajira Department, Colombia. The full translation is 'titanic boa from Cerrejón'.

Holotype. UF/IGM 1, a single precloacal vertebra (Fig. 1a–d).

Locality. La Puente Pit, Cerrejón Coal Mine, Guajira Peninsula, Colombia (palaeolatitude 5.5° N; Supplementary Fig. 1).

Horizon. Single claystone layer, middle segment of the Cerrejón Formation (Supplementary Fig. 2); middle–late Palaeocene epoch (58–60 Myr ago), palynological zone Cu-02 (ref. 11).

Referred material. UF/IGM 2 (paratype), nearly complete precloacal vertebra (Fig. 1g, h). UF/IGM 3–UF/IGM 28, 184 additional precloacal vertebrae and ribs representing 28 individuals (Supplementary Table 1).

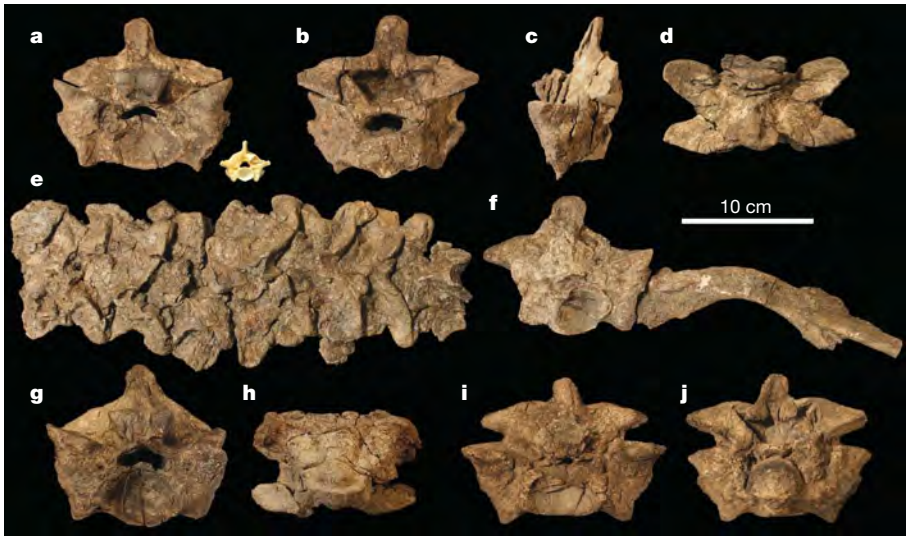


Figure 1 | *Titanoboa cerrejonensis* precloacal vertebrae. **a**, Type specimen (UF/IGM 1) in anterior view compared to scale with a precloacal vertebra from approximately 65% along the precloacal column of a 3.4 m *Boa constrictor*. Type specimen (UF/IGM 1) shown in posterior view (**b**), left lateral view (**c**) and dorsal view (**d**). Seven articulated precloacal vertebrae (UF/IGM 3) in dorsal view (**e**). Articulated precloacal vertebra and rib (UF/IGM 4) in anterior view (**f**). Precloacal vertebra (paratype specimen UF/IGM 2) in anterior view (**g**) and ventral view (**h**). Precloacal vertebra (UF/IGM 5) in anterior view (**i**) and posterior view (**j**). All specimens are to scale.

[1]Department of Biology, University of Toronto, Mississauga, Ontario L5L 1C6, Canada. [2]Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611-7800, USA. [3]Smithsonian Tropical Research Institute, Box 0843-03092, Balboa, Ancon Republic of Panama. [4]Department of Geological Sciences, Indiana University, Bloomington, Indiana 47405-1405, USA.

Diagnosis. Extremely large-bodied boine snake with robust precloacal vertebrae possessing a uniquely T-shaped neural spine composed of a transversely expanded posterior margin and distinctly narrow, blade-like anterior process (Fig. 1a–d, i, j). Subcentral and lateral foramina are extremely reduced.

The vertebrae possess a character combination unique to boine snakes. These are: the presence of paracotylar fossae and foramina; straight, posteromedially angled interzygapophyseal ridges; and the vaulted, bi-angled posterior margin of the neural arch. These characters are also present in some madtsoiid snakes; however, all specimens of *Titanoboa* possess short, posteriorly angled prezygapophyseal accessory processes as in boines but unlike madtsoiids, and lack the parazygantral foramina and laterally extensive synapophyses that diagnose Madtsoiidae[12]. Among extant boines, *Titanoboa* is united with *Boa constrictor* on the basis of dorsolaterally positioned paracotylar fossae and foramina.

Vertebrae of *Titanoboa* are the largest recovered for any extant or fossil snake[6–8]. Body size can be predicted from vertebral dimensions in taxa where body length evolved by increasing the size of vertebrae instead of their number. This is true for all extant giant boids and pythonids[13] and is inferred for *Titanoboa* because it is united with *Boa* within Boinae. Vertebral size changes along the vertebral column in snakes, and the position of isolated fossil vertebrae, must be determined before body length can be reconstructed. We estimated vertebral position by matching the vertebral shape of two undistorted specimens of *Titanoboa* to a composite geometric morphometric model vertebral column[14] constructed from extant boines (see Methods). Both vertebrae were estimated to be located 60–65% back along the precloacal vertebral column from the axis–atlas complex. Regressions of vertebral width from this region against body lengths for extant boines indicate a snout–vent length (SVL) of 12.01 ± 2.04 m (39 ft) and a total body length (TBL) of 12.82 ± 2.18 m (42 ft) for *Titanoboa*. Incorporating SVL values of this study into the relationship between length and body mass determined for extant *Eunectes murinus* (green anaconda)[2] and *Python natalensis* (southern African python)[15] results in an estimated mass for *Titanoboa* of 1,135 kg (1.27 ton) with a range of 652–1,819 kg (0.73–2.03 ton).

Body size estimates for *Titanoboa* greatly exceed the largest verifiable body lengths for extant *Python* and *Eunectes*, which are approximately 9 m and 7 m, respectively[1]. Maxima for these taxa are extraordinary, however, and surveys of large populations have not recovered individuals exceeding 6 m TBL for *Python* and 6.5 m TBL for *Eunectes*[2,3,15,16]. Conversely, the record of *Titanoboa* includes eight individuals represented by vertebrae of approximately the same size as the elements used to estimate TBL (Fig. 1, Supplementary Table 1), indicating that extremely large body size was common in the taxon. *Titanoboa* is larger than all other giant fossil taxa, including palaeopheids and madtsoiids[6,9], making it the largest known snake (Fig. 2). Discovery of *Titanoboa* extends the known range of body lengths in snakes by more than two orders of magnitude, between TBLs of 10 cm (*Leptotyphlops carlae*) and 12.8 m. Our estimates of body size also demonstrate that *Titanoboa* is the largest known non-marine vertebrate from the Palaeocene and early Eocene[17].

Large body size in *Titanoboa* provides significant information on equatorial climates during the Palaeogene. Snakes have body temperatures that are dependant on their ambient environment (poikilothermy), and ambient temperature regulates maximum body size in poikilothermic vertebrates[4,5]. Palaeotemperature can be predicted from fossils of poikilothermic taxa using a model for extant taxa[4] that demonstrates that the difference in maximum body size of taxa between two localities is proportional to the difference in ambient temperature for a given mass-specific metabolic rate (see Methods). We used the difference in TBL between *Titanoboa* and *Eunectes murinus*, the largest snake in the modern neotropics, to reconstruct the mean annual temperature (MAT) for the Palaeocene of equatorial South America. The relationship between TBL and temperature in *Eunectes* indicates that the approximate minimum MAT under which a 13-m-long boine snake could survive is 32–33 °C, ranging between 30 °C and 34 °C for body sizes between 11 m and 15 m (Fig. 3).

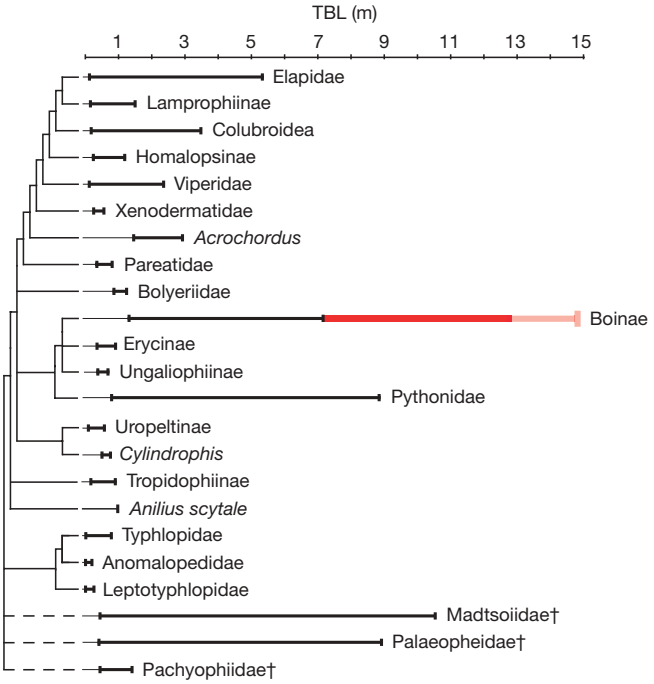


Figure 2 | Body size ranges for major snake clades plotted along phylogeny[28–30] (Supplementary Table 3). Controversial fossil (dagger) lineages Madtsoiidae, Pachyophiidae and Palaeopheidae were placed as an unresolved polytomy at the base of the snake crown. The size range increase in Boinae based on the *Titanoboa cerrejonensis* mean TBL estimate is in dark red; maximum TBL estimate for *Titanoboa* is in pink.

These temperature estimates are consistent with hot Palaeogene climate models requiring high atmospheric p_{CO2} concentrations of approximately 2,000 parts per million[18], and are slightly higher than temperatures derived from planktonic foraminifer oxygen isotopes by 1–5 °C[19]. These estimates exceed MATs derived from coeval Cerrejón palaeofloras by 6–8 °C[20], but palaeotemperatures based on fossil leaf assemblages from riparian and wetland habitats of rainforests are underestimates[21]. Palaeotemperature estimates of 30–34 °C exceed MAT maxima of modern tropical forests[22]. However, the high rainfall estimates from the Cerrejón palaeoflora (~4 m per year[11]) combined with increased p_{CO_2} could have maintained forest floras under higher temperature conditions[23].

Palaeotemperature estimates near the equator allow reconstruction of latitudinal temperature gradients across South America during the Palaeogene. MAT for the middle Palaeocene of Argentina (palaeolatitude ~51° S) is 14.1 °C ± 2.6 °C[24], indicating a latitudinal gradient of 13–22 °C between 5° N and 51° S, with a midpoint of 18 °C (accounting for taphonomic bias[21] suggests MAT of 17.6 °C ± 3.6 °C with a gradient midpoint of 15.4 °C). Our midpoint estimates during the early Palaeogene greenhouse approximate the modern temperature difference across South America (Fig. 3) and are not consistent with the climatic thermostat hypothesis that predicts cooler equatorial temperatures and a shallow temperature gradient during greenhouse intervals[25]. If our Palaeocene estimates are correct, tropical temperatures at the slightly younger (55.8 Myr ago) Palaeocene–Eocene thermal maximum (PETM) could have reached 38–40 °C, resulting in widespread equatorial heat-death as recent models and other proxy data have predicted[26]. However, we still lack empirical evidence of the effects of the PETM on tropical floras and faunas.

Remains of *Titanoboa* were found in depositional environments consisting of coastal plains incised by large-scale river systems within a wet tropical rainforest[11,20] and were associated with an aquatic vertebrate fauna including podocnemidid pleurodire turtles, dyrosaurid

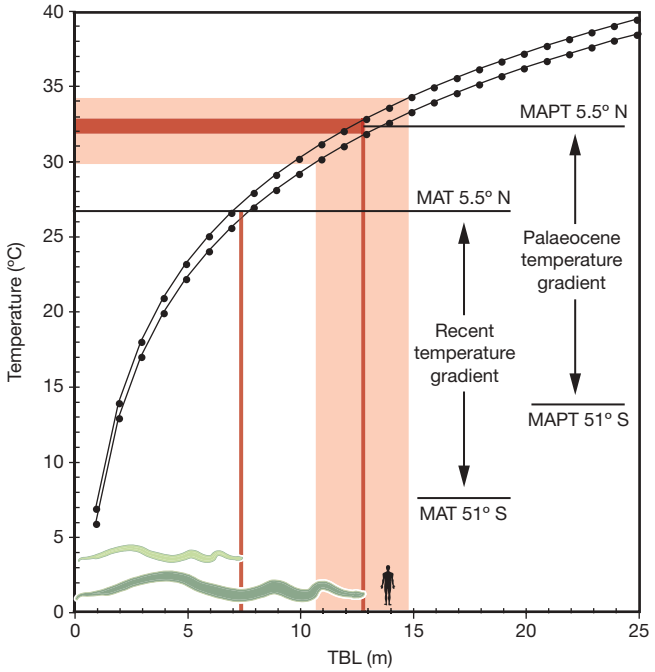


Figure 3 | Mean annual palaeotemperature and Palaeocene latitudinal temperature gradients derived from body size of the green anaconda *Eunectes murinus* (light green) and body size estimates of *Titanoboa cerrejonensis* (dark green). Curves represent model body size increases with temperature in boine snakes based on a maximum TBL for *Eunectes* of 7.3 m at modern neotropical MAT of 26 °C (lower curve) and 27 °C (upper curve). Light red regions indicate error for *Titanoboa* TBLs and resultant temperature ranges. A MAPT gradient of 18 °C from equatorial to mid-latitudes at ~58 Myr ago is equivalent to the modern gradient (18–19 °C). Silhouettes are to scale for *Titanoboa*, *Eunectes* and a 1.85-m-tall adult human male.

mesoeucrocodylians, and elopomorph and dipnoan fishes. Similarities between depositional environments of the Cerrejón Formation and habitats of extant *Eunectes* together with inferred prey taxa (crocodyliforms) indicate a similar ecology of *Titanoboa* to modern anacondas[2,3]. Discovery of *Titanoboa* and the additional Cerrejón Formation fossil record indicates that components of modern neotropical riverine vertebrate faunas were assembled at most six to seven million years after the Cretaceous–Palaeogene extinction event.

METHODS SUMMARY

We estimated SVL and TBL in *Titanoboa* by first determining the intracolumnar position of isolated precloacal vertebrae through maximum likelihood identification of quantified vertebral morphology against morphological change along a model boine vertebral column. We regressed SVL and TBL of extant taxa onto vertebral width (postzygapophyseal width) for the intracolumnar regions corresponding to the positions determined for the fossil elements, and used the resulting equations to calculate SVL and TBL for fossil specimens. We estimated mean annual palaeotemperatures (MAPT) by solving the equation describing size differences across a temperature gradient at a standard coefficient of metabolism [Q_{10}][27], TBLs for *Titanoboa* of 10.6–14.9 m, maximum TBL for *Eunectes murinus* of 7.3 m[1], and MAT values for modern neotropical lowland rainforests of 26–27 °C[22].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 11 October; accepted 26 November 2008.

1. Murphy, J. C. & Henderson, R. W. *Tales of Giant Snakes: A Natural History of Anacondas and Pythons* (Krieger, 1997).
2. Rivas, J. *The Life History of the Green Anaconda (Eunectes murinus), with Emphasis on its Reproductive Biology*. Dissertation, Univ. Tennessee (1999).
3. Dirksen, L. *Anakondas: monographische Revision der Gattung Eunectes Wagler 1830 (Serpentes, Boidae)* (Natur und Tier, 2002).
4. Makarieva, A. M., Gorshkov, V. G. & Li, B.-L. Gigantism, temperature and metabolic rate in terrestrial poikilotherms. *Proc. R. Soc. Lond. B* **272,** 2325–2328 (2005).
5. Makarieva, A. M., Gorshkov, V. G. & Li, B.-L. Temperature-associated upper limits to body size in terrestrial poikilotherms. *Oikos* **111,** 425–436 (2005).
6. Rage, J.-C. *Palaeophis colossaeus* nov. sp. (le plus grand Serpent connu?) de l'Eocène du Mali et le problème du genre chez les Palaeopheinae. *C.R. Acad. Sci. Sér.* **2,** 1741–1744 (1983).
7. Albino, A. M. Serpientes gigantes en la Patagonia. *Ciencia Hoy* **3,** 58–63 (1991).
8. Scanlon, J. D. & Mackness, B. S. A new giant python from the Pliocene Bluff Downs local fauna of northeastern Queensland. *Alcheringa* **25,** 425–437 (2002).
9. Head, J. J. & Polly, P. D. They might be giants: morphometric methods for reconstructing body size for the World's largest snakes. *J. Vertebr. Paleontol.* **24** (suppl. 3), 68A (2004).
10. Sloan, L. C. & Shellito, L. J. in *Causes and Consequences of Globally Warm Climates in the Early Paleogene* (eds Wing, S. L., Gingerich, P. D., Schmitz, B. & Thomas, E.) 25–47 (Geological Society of America Special Paper, 369, 2003).
11. Jaramillo, C. *et al.* Palynology of the upper Paleocene Cerrejón Formation, Northern Colombia. *Palynology* **31,** 153–189 (2007).
12. Scanlon, J. D. Skull of the large non-macrostomatan snake *Yurlunggur* from the Australian Oligo–Miocene. *Nature* **439,** 839–842 (2006).
13. Head, J. J. & Polly, P. D. Dissociation of somatic maturity from segmentation drives gigantism in snakes. *Biol. Lett.* **3,** 296–298 (2007).
14. Polly, P. D. & Head, J. J. in *Morphometrics-Applications in Biology and Paleontology* (ed. Elewa, A. M. T.) 197–222 (Springer, 2004).
15. Alexander, G. J. in *Biology of the Boas and Pythons* (eds Henderson, R. W. & Powell, R.) 51–75 (Eagle Mountain Publishing, 2007).
16. Shine, R., Harlow, P. S., Keogh, J. S. & Boeadi. The influence of sex and body size on food habits of a giant tropical snake, *Python reticulatus. Funct. Ecol.* **12,** 248–258 (1998).
17. Alroy, J. Cope's rule and the dynamics of body mass evolution in North American fossil mammals. *Science* **280,** 731–734 (1998).
18. Shellito, C. J., Sloan, L. C. & Huber, M. Climate model sensitivity to atmospheric CO_2 levels in the Early–Middle Paleogene. *Palaeogeogr. Palaeclimatol. Palaeoecol.* **193,** 113–123 (2003).
19. Pearson, P. N. *et al.* Stable warm tropical climate through the Eocene epoch. *Geology* **35,** 211–214 (2007).
20. Herrera, F., Wing, S. & Jaramillo, C. Warm (not hot) tropics during the Late Paleocene. First Continental Evidence. *Eos Trans. AGU* **86** (Suppl.), PP51C–0608 (2005).
21. Kowalski, E. A. & Dilcher, D. L. Warmer paleotemperatures for terrestrial ecosystems. *Proc. Natl Acad. Sci. USA* **100,** 167–170 (2003).
22. Burnham, R. J. & Johnson, K. R. South American paleobotany and the origins of neotropical rainforests. *Phil. Trans. R. Soc. Lond. B* **359,** 1595–1610 (2004).
23. Hogan, K. P., Smith, A. P. & Ziska, L. H. Potential effects of elevated CO_2 and changes in temperature on tropical plants. *Plant Cell Environ.* **14,** 763–778 (1991).
24. Iglesias, A. *et al.* A Paleocene lowland macroflora from Patagonia reveals significantly greater richness than North American analogs. *Geology* **35,** 947–950 (2007).
25. Crowley, T. J. & Zachos, J. C. in *Warm Climates in Earth History* (eds Huber, B. T., MacLeod, K. G. & Wing, S. L.) 50–76 (Cambridge Univ. Press, 2000).
26. Huber, M. A hotter greenhouse? *Science* **321,** 353–354 (2008).
27. Chappell, M. A. & Ellis, T. M. Resting metabolic rates in boid snakes: allometric relationships and temperature effects. *J. Comp. Physiol. B* **157,** 227–235 (1987).
28. Vidal, N. & Hedges, S. B. Higher-level relationships of snakes inferred from four nuclear and mitochondrial genes. *C.R. Biol.* **325,** 977–985 (2002).
29. Lawson, R., Slowinski, J. B. & Burbrink, F. T. A molecular approach to discerning the phylogenetic placement of the enigmatic snake *Xenophidion schaeferi* among the Alethinophidia. *J. Zool. (Lond.)* **263,** 285–294 (2004).
30. Vidal, N. *et al.* The phylogeny and classification of caenophidian snakes inferred from seven nuclear protein-coding genes. *C.R. Biol.* **330,** 182–187 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Bell, R. Ghent, E. Kowalski, A. M. Lawing, B. MacFadden, R. Reisz and S. Wing for advice and discussion, K. Seymour, K. Krysko, K. deQueiroz and G. Zug for access to comparative specimens, A. Rincon and M. Carvalho for fieldwork, J. Mason, K. Church, J. Mathis and J. Nestler for fossil preparation, and K. Krysko and J. Nestler for photographic assistance. We thank Carbones del Cerrejón, L. Teicher, F. Chavez, C. Montes and G. Hernandez for logistical support and access to the Cerrejón mine. This research was funded by the National Science Foundation, Fondo para Investigaciones del Banco de la Republica de Colombia, Smithsonian Tropical Research Institute Paleobiology Fund, the Florida Museum of Natural History, a Geological Society of America Graduate Student Research Grant to A.K.H., and a National Sciences and Engineering Research Council of Canada Discovery Grant to J.J.H.

Author Contributions J.J.H., J.I.B., C.A.J., P.D.P., A.K.H. and J.R.B. contributed to project planning. J.J.H. and J.I.B. contributed to systematic palaeontology. J.J.H., P.D.P., J.I.B., A.K.H., J.R.B. and E.A.C. contributed to body size estimation. J.J.H., J.I.B., F.A.H., P.D.P. and C.A.J. contributed to palaeoclimatic analysis. J.I.B., A.K.H., E.A.C., F.A.H. and C.A.J. contributed to fieldwork. J.I.B., A.K.H., C.A.J. and J.J.H. contributed to financial support. All authors contributed to manuscript and figure preparation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.J.H. (jason.head@utoronto.ca).

METHODS

Position estimation of fossil vertebrae. We used a maximum-likelihood algorithm to find the most likely position of isolated vertebrae along an anterior–posterior shape gradient derived from the vertebrae of representative extant taxa, a procedure modified from ref. 14. The anterior–posterior morphological gradient was estimated by measuring the shape of vertebrae between the first and last precloacal vertebrae. Vertebral shape was quantified using two-dimensional geometric landmarks that represent morphology in anterior view (Supplementary Fig. 3). The anterior view was chosen because it provides height and width information as well as the most complex vertebral morphology. The number of precloacal vertebrae varies within Boinae[13] (Supplementary Table 2), so we sampled vertebral morphology at 5% intervals along the column for all specimens to standardize comparisons across taxa.

We projected the vertebral landmarks of all the extant species and the isolated fossil specimens into the same shape space by Procrustes superimposition to minimize shape differences among specimens, with orthogonal projection into tangent space. Shapes were rotated to their principal components (PC) axes using singular value decomposition to find the eigenvectors and eigenvalues. The PC axes have the valuable property that the shape variation described by each one is statistically uncorrelated with the shape variation described by the others. Variance is therefore additive across the axes, allowing the PC scores to be used as uncorrelated variables in multivariate statistical analysis.

A multivariate regression of vertebral shape onto position in the vertebral column was used to extract a species-independent anterior–posterior shape gradient from the extant data set. Both a discrete function and a continuous spline function were fit. The discrete function runs through the multivariate means of each of the 5% vertebral positions and is undefined between them. The spline function runs through the 5% multivariate means and is interpolated between them (Supplementary Fig. 4).

These multivariate regression functions and their residual variation were used as likelihood models for estimating the position of the fossil vertebrae. The following function describes the likelihood distribution of shape at vertebral position (pos) k:

$$L(\mathrm{pos}_k|z,\hat{z},\sigma^2)=\prod_{n=1}^{i}\frac{1}{\sqrt{2\pi{\sigma_{k,i}}^2}}e^{-\frac{(z_i-\hat{z}_{k,i})^2}{2{\sigma_{k,i}}^2}} \quad (1)$$

where i is the number of principal components, k,i is the score of the unknown vertebra on PC_i, $\hat{z}_{k,i}$ is the expectation of shape at vertebral position k on PC_i along the shape gradient defined by the extant species, and ${\sigma^2}_i$ is the residual variance around the estimated shape gradient at position k on PC_i. If the variance is presumed to be equal along the length of the shape gradient, which it is approximately in our data, then the variance term becomes a constant and can be dropped:

$$L(\mathrm{pos}_k|z,\hat{z})=\prod_{n=1}^{i}e^{-(z_i-\hat{z}_{k,i})^2} \quad (2)$$

The log likelihood equation is then simply:

$$l(\mathrm{pos}_k|z,\hat{z})=\sum_{n=1}^{i}(z_i-\hat{z}_{k,i})^2 \quad (3)$$

Maximizing this equation for position k gives the best estimate of the position of the unknown vertebrae given its shape (z) and the estimated shape gradient of the extant snakes ($\hat{z}$).

Standard errors (s.e.) for the positional estimates were obtained by cross validation. Isolated vertebrae with a known position were systematically selected from the sample of extant snakes. Each vertebra was submitted to the maximum likelihood procedure and the distance of the estimated position from the true position was noted. s.e. is the mean distance from the true value for the entire sample.

Regression of body size onto vertebral size. We estimated body length for *Titanoboa* by regressing SVL and TBL measured in millimetres onto postzygapophyseal width measured in millimetres for precloacal vertebrae between 60% and 65% intervals along the precloacal vertebral column for the examined sample of extant boines ($n = 21$, Supplementary Table 2), based on results of position estimation, and applying the resultant equation to the holotype (UF/IGM 1, width = 120 mm) and paratype (UF/IGM 2, width = 119 mm) specimens. SVL, TBL and vertebral width data were not log-transformed because they were approximately normally distributed (SVL skewness = 0.63, TBL skewness = 0.49, postzygapophyseal width skewness = 0.64, s.e. skewness for all = 1.07). Least-squares linear regression models produced positive, significant relationships between SVL and width (60%: slope = 95.9, intercept = 262.6, $P < 0.001$, $R^2 = 0.85$; 65%: slope = 100.4, intercept = 226.5, $P < 0.001$, $R^2 = 0.87$), and TBL and width (60%: slope = 100.7, intercept = 436.2, $P < 0.001$, $R^2 = 0.81$; 65%: slope = 106.0, intercept = 390.0, $P < 0.001$, $R^2 = 0.83$). The estimated means of 12.04 m SVL and 12.82 m TBL were obtained by averaging the 60% and 65% estimates. The error for size estimates was determined by subtracting the averaged regression coefficients from a perfect fit for extant taxa.

Palaeotemperature estimation. We estimated palaeotemperature from body size using the equation of ref. 4:

$$\frac{L_1}{L_2}=Q_{10}^{(\Delta T/10^{\circ}\mathrm{C})/3\alpha} \quad (4)$$

where L_1 is length of the largest taxon, L_2 is length of the smallest taxon, 10 °C is interval of temperature change associated with metabolic rate change (Q_{10}; ref. 5), and $\Delta T = \mathrm{temperature}_1 - \mathrm{temperature}_2$. We solved for the temperature associated with the larger taxon (*Titanoboa*; Fig. 3) as follows:

$$\mathrm{MAPT}=\mathrm{MAT}+3\alpha 10^{\circ}\mathrm{C}\left(\frac{\log_{10}(\mathrm{TBL_T}/\mathrm{TBL_E})}{\log_{10}Q_{10}}\right) \quad (5)$$

where MAPT is mean annual palaeotemperature ($\mathrm{temperature}_1$ in equation (4)), MAT is modern mean annual temperature ($\mathrm{temperature}_2$ of equation (4)), $\mathrm{TBL_T}$ is total body length of *Titanoboa* (L_1 in equation (4)), $\mathrm{TBL_E}$ is total body length of *Eunectes* (L_2 of equation (4)), Q_{10} is mass-specific metabolic rate of 2.65 for boid snakes[27], and $\alpha = 0.33$ (ref. 5):

$$\mathrm{MAPT}=\mathrm{MAT}+9.9^{\circ}\mathrm{C}\left(\frac{\log_{10}(\mathrm{TBL_T}/\mathrm{TBL_E})}{0.42}\right) \quad (6)$$

Vol 457|5 February 2009|doi:10.1038/nature07673

LETTERS

Fossil steroids record the appearance of Demospongiae during the Cryogenian period

Gordon D. Love[1,2], Emmanuelle Grosjean[3], Charlotte Stalvies[4], David A. Fike[5], John P. Grotzinger[5], Alexander S. Bradley[2], Amy E. Kelly[2], Maya Bhatia[2], William Meredith[6], Colin E. Snape[6], Samuel A. Bowring[2], Daniel J. Condon[2]† & Roger E. Summons[2]

The Neoproterozoic era (1,000–542 Myr ago) was an era of climatic extremes and biological evolutionary developments culminating in the emergence of animals (Metazoa) and new ecosystems[1]. Here we show that abundant sedimentary 24-isopropylcholestanes, the hydrocarbon remains of C_{30} sterols produced by marine demosponges, record the presence of Metazoa in the geological record before the end of the Marinoan glaciation (~635 Myr ago). These sterane biomarkers are abundant in all formations of the Huqf Supergroup, South Oman Salt Basin, and, based on a new high-precision geochronology[2], constitute a continuous 100-Myr-long chemical fossil record of demosponges through the terminal Neoproterozoic and into the Early Cambrian epoch. The demosponge steranes occur in strata that underlie the Marinoan cap carbonate (>635 Myr ago). They currently represent the oldest evidence for animals in the fossil record, and are evidence for animals pre-dating the termination of the Marinoan glaciation. This suggests that shallow shelf waters in some late Cryogenian ocean basins (>635 Myr ago) contained dissolved oxygen in concentrations sufficient to support basal metazoan life at least 100 Myr before the rapid diversification of bilaterians during the Cambrian explosion. Biomarker analysis has yet to reveal any convincing evidence for ancient sponges pre-dating the first globally extensive Neoproterozoic glacial episode (the Sturtian, ~713 Myr ago in Oman[2]).

The Neoproterozoic-Cambrian Huqf Supergroup, South Oman Salt Basin (SOSB), is located at the southeastern edge of the Arabian peninsula and comprises the Abu Mahara Group encompassing Sturtian- and Marinoan-equivalent glacial deposits, and the Nafun and Ara Groups[2,3] (Fig. 1). The Abu Mahara Group was deposited in localized rift basins, whereas the Nafun Group records two shallowing-upward siliciclastic-carbonate sequences (Masirah Bay Formation–Khufai Formation; Shuram Formation–Buah Formation) deposited in a regionally extensive sag basin[4]. The Ara Group, which was deposited ~547–540 Myr ago[2], consists of a series of carbonate-evaporite sequences (A0–A6) within the SOSB preserved solely in the subsurface. The Ara Group contains the Ediacaran–Cambrian boundary at the base of the fourth (A4) carbonate unit. Well-preserved lipid biomarkers are prevalent in the sedimentary rocks and oils of the Huqf. Previous organic geochemical studies show that SOSB oils, and their precursor source rocks, have a very distinctive molecular and isotopic geochemistry marked by unusual abundances of methylalkanes, steroids and triterpenoids derived from microbiota present at the time of sediment deposition[5,6].

We analysed extractable saturated and aromatic hydrocarbons from 64 sedimentary rock samples, comprising core and cuttings, from 26 different wells from the petroleum-rich SOSB (Fig. 1). Analyses were carried out via gas chromatography (GC) and gas chromatography-mass spectrometry (GC-MS) with the high sensitivity, selectivity and accuracy afforded by multiple-reaction-monitoring (MRM) mass spectrometry (see Supplementary Information). To establish the stratigraphic range of specific organic compounds beyond doubt, we isolated kerogens (insoluble, macromolecular organic matter that cannot migrate) from key samples. From these kerogens we generated complementary sets of biomarkers using catalytic hydropyrolysis (HyPy). With this technique, covalently bound hydrocarbons are released from the (immobile) kerogen by continuous-flow, temperature-programmed pyrolysis in a stream of high-pressure (15 MPa) H_2 gas and using a molybdenum sulphide catalyst. HyPy is a powerful analytical tool for obtaining high yields of biomarker hydrocarbons with optimal preservation of structure and stereochemistry[7]. Kerogen-bound biomarkers released by HyPy can be unambiguously correlated to a specific stratigraphic interval.

The absolute abundances of extractable C_{26}–C_{30} steranes (which ranged from ~300 to 13,000 p.p.m. of total saturated hydrocarbons, depending on thermal maturity) and sterane/hopane ratios (0.21–1.50, with an average value of 0.81; Table 1 and Supplementary Table 1) in these Huqf samples are comparable in magnitude to those found in typical Phanerozoic marine organic matter such as the Kimmeridge Clay[8] that sources North Sea petroleum. This contrasts with the trace amounts of regular steranes detected (<1 p.p.m. of total organic carbon) in rock extracts of similar thermal maturity from highly euxinic facies of the 1,640-Myr-old Barney Creek Formation[9]. Eukaryotic microalgae are most probably the principal biological source of steranes in Neoproterozoic-Cambrian sedimentary rocks. Their high absolute concentrations in Huqf sedimentary rocks suggests that marine microbial communities rich in microalgae proliferated in Neoproterozoic oceans.

Accumulation of abundant hopanes and 2-methylhopanes in Huqf sedimentary rocks suggests that bacteria[10] constituted a significant fraction of primary productivity, but the absolute abundance of C_{29} steranes and their dominance over C_{26}–C_{30} steranes (Table 1 and Supplementary Tables 1 and 2) suggests that chlorophyte microalgae were quantitatively important as marine primary producers. This sterane pattern mirrors the C_{29} sterol carbon number dominance in many extant chlorophytes[11]. The prominence of C_{29} steranes over other steranes is observed in all SOSB formations, including the Cryogenian Ghadir Manquil Formation. High diversity in the structures of the minor SOSB steranes indicates that other groups of microalgae must also have been present, including marine pelagophytes

[1]Department of Earth Sciences, University of California, Riverside, California 92521, USA. [2]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 01239, USA. [3]Petroleum and Marine Division, Geoscience Australia, Canberra, Australian Capital Territory 2601, Australia. [4]School of Civil Engineering and Geosciences, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, UK. [5]Department of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA. [6]School of Chemical, Environmental and Mining Engineering, University of Nottingham, University Park, Nottingham NG7 2RD, UK. †Present address: NERC Isotope Geosciences Laboratory, Keyworth, Nottingham NG12 5GG, UK.

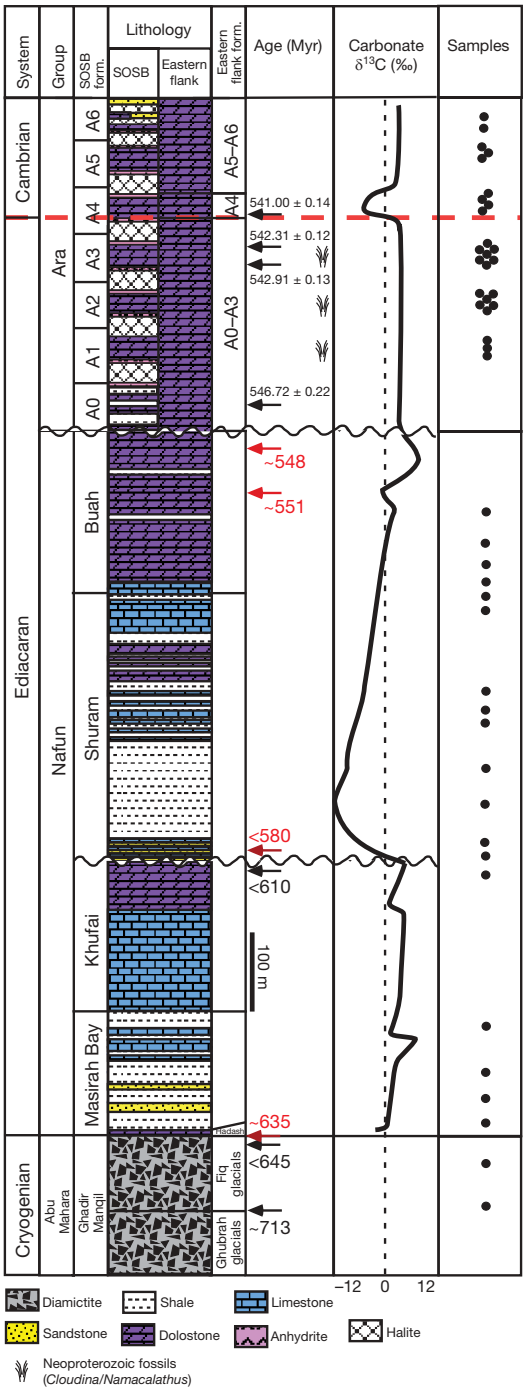


Figure 1 | Stratigraphic column of Huqf Supergroup with representative lithology, biostratigraphy and geochronological constraints. Stratigraphic distribution of samples in the present study is indicated on the right. Absolute dates in red are from correlation with other dated sections worldwide (Namibia, South China) using comparisons of $\delta^{13}C$ carbonate stratigraphic features. Absolute dates[2] in black are from direct U–Pb zircon age measurements on Huqf detrital zircons and ash beds. See Supplementary Information for additional discussion. A typical $\delta^{13}C$ carbonate stratigraphic profile is drawn for reference[3].

and dinoflagellates inferred from 24-*n*-propyl steranes[12] and dinosteranes[11], respectively. Uncommon steranes detected in these rocks included 27-norcholestanes, 21-norcholestanes, 21-norergostanes and 21-norstigmastanes and a variety of C_{19} to C_{20} steroids that have had their side-chains excised. Of particular note was the high relative abundance of C_{30} steranes with a 24-isopropyl moiety in all formations of the Huqf Supergroup (Fig. 2, Table 1, Supplementary Tables 1 and 2) which signifies demosponge inputs.

Table 1 | Summary of key biomarker ratios for Huqf rock bitumens and kerogen hydropyrolysates

Formation	Phase	Ster/hop*	%C_{29} ster†	%C_{30} ster‡	*i*-C_{30}/*n*-C_{30}§
Ara carbonates	Bitumen [25]	0.2–1.1 (0.8)	52–75 (69)	1.8–6.7 (2.7)	1.0–1.9 (1.5)
Ara carbonates	Kerogen [8]	0.6–1.0 (0.8)	52–70 (61)	1.9–3.6 (2.8)	0.6–1.4 (0.9)
Thuleilat‖	Bitumen [5]	0.6–1.3 (0.9)	60–73 (68)	1.9–2.8 (2.5)	1.3–1.6 (1.4)
Thuleilat	Kerogen [2]	1.0–2.5	57–65	2.0–2.5	0.7–1.2
Silicylite‖	Bitumen [5]	0.8–1.5 (1.1)	72–76 (74)	1.8–2.4 (2.1)	1.4–2.4 (1.8)
Silicylite	Kerogen [2]	1.5–2.1	69–75	2.1–2.3	0.7–1.1
U shale‖	Bitumen [5]	0.8–1.0 (0.9)	60–66 (63)	2.2–2.9 (2.5)	0.8–1.4 (1.1)
U shale	Kerogen [1]	1.2	53	2.7	1.7
Buah	Bitumen [4]	0.7–1.1 (0.9)	64–73 (69)	1.3–1.8 (1.6)	0.6–0.9 (0.7)
Buah	Kerogen [2]	0.9–1.1	63–67	1.3–1.9	0.5–0.7
Shuram	Bitumen [8]	0.6–1.1 (0.8)	65–77 (70)	1.9–2.5 (2.2)	0.8–1.8 (1.2)
Shuram	Kerogen [2]	0.6–1.0	58–60	2.7–4.6	0.7–1.4
Khufai	Bitumen [2]	0.5–0.8	72–73	2.0–3.3	1.3–1.4
Masir. B.	Bitumen [5]	0.5–0.7 (0.7)	58–84 (67)	2.3–13 (4.8)	1.3–16 (4.9)
Masir. B.	Kerogen [2]	0.7–1.2	54–55	3.4–5.6	1.4–1.5
Gh. Manq.	Bitumen [2]	0.4–0.9	56–72	2.7–3.7	0.5–3.3
Gh. Manq.	Kerogen [1]	0.7	66	3.0	1.3

[*n*] represents number of samples; () are average ratio values for $n > 2$; a more comprehensive biomarker data set is given in Supplementary Tables 1 and 2. Average uncertainties in hopane and sterane biomarker ratios are ±8% as calculated from multiple analyses of saturated hydrocarbon fractions prepared from an AGSO standard oil ($n = 30$). Masir. B., Masirah Bay; Gh. Manq., Ghadir Manquil.
* Ratio of (C_{27}–C_{29} steranes)/(C_{27} + C_{29-35} hopanes).
† Ratio of C_{29} steranes to ΣC_{27}–C_{29} steranes.
‡ (24-*n*-propylcholestanes + 24-isopropylcholestanes)/Σ(C_{27}–C_{30} steranes).
§ 24-*i*-propylcholestanes/24-*n*-propylcholestanes (using all 4 regular isomers).
‖ Formations in the basin centre (Athel basin) which are age equivalent to Ara Group carbonates (see Supplementary Information).

24-Isopropylcholestane is the geologically stable form of 24-isopropylcholesterol and related structures, which are primarily found in certain genera of the Demospongiae[13] and can be biosynthesized *de novo* to function in the sponge cell membrane[14] (Supplementary Information). 24-Isopropylcholestanes were previously shown to be abundant relative to microalgal 24-*n*-propylcholestanes (>0.5) in numerous Ediacaran to Early Cambrian oils and calcareous sediments, thus representing anomalously elevated levels of these compounds (Supplementary Information), and on this basis, were proposed as molecular fossils of sponges or their ancestors[15]. Potential precursor sterols were not identified in the choanoflagellate *Monosiga brevicollis*[16], a representative for the unicellular sister group of animals. The rigorous stratigraphic and geochronologic placement of the SOSB samples in our study constrains the first appearance of sponge biomarkers and suggests that sponges were continuously prevalent in a wide range of Neoproterozoic environments before the known record of other animal fossils, including megascopic animal body fossils ~575 Myr ago[17], trace fossils ~555 Myr ago[18] and putative animal embryos <632 to >550 Myr ago[19]. The detection of free and kerogen-bound sponge steranes in sedimentary rocks from the Ghadir Manquil Formation (Fig. 2) of the Huqf Supergroup, found stratigraphically below the Marinoan cap carbonate, suggests a Cryogenian origin of Metazoa. Detrital zircon U–Pb ages of ~751 Myr were obtained previously from Ghadir Manquil Formation sediments from SOSB[2], including from the GM-1 well (Supplementary Fig. 1), so 751 Myr constitutes a maximum age for the Cryogenian sponge biomarkers found in our study. Analysis of a number of pre-Sturtian sediments from other sections worldwide has found no convincing evidence for elevated levels of 24-isopropylcholestanes in rock bitumens (Supplementary Information).

Existing fossil evidence for Ediacaran sponges comes from detection of siliceous spicules derived from hexactinellids in ~543–549 Myr sedimentary rocks from Australia[20] and southwestern Mongolia[21], and from putative siliceous demosponge spicules[22] found alongside preserved sponge tissue and animal embryos[23] in <600-Myr Doushantuo phosphorites in South China. Molecular phylogenetic classifications using metazoan protein amino acid and nucleic acid sequences usually place the silicisponges, the demosponges and hexactinellids, as the earliest diverging animals[24]. The timing of the sponge

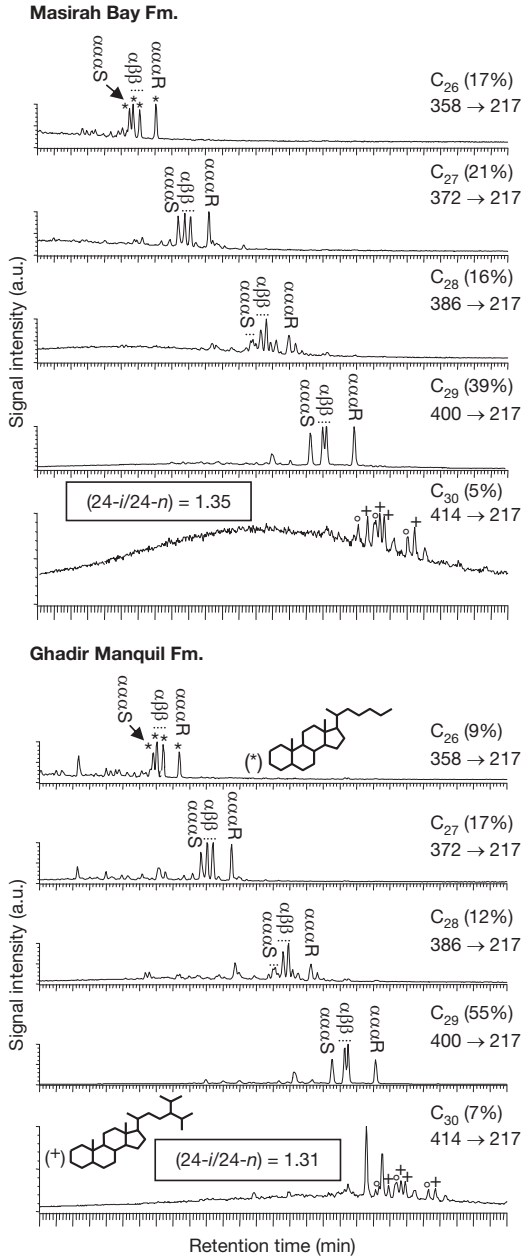


Figure 2 | MRM GC-MS ion chromatograms of C_{26}–C_{30} desmethylsteranes released from catalytic hydropyrolysis of a Masirah Bay Formation (JF-1) and a Ghadir Manquil Formation (GM-1) kerogen. For each sterane carbon number, four diastereoisomers are detected (ααα20R, αββ20R, αββ20S, ααα20R), indicating a mature geoisomer distribution. Demosponge contributions are evident from abundant 24-isopropylcholestanes ('plus' signs). 24-*n*-propylcholestanes (open circles) are markers of marine pelagophyte algae and this confirms a marine depositional setting for each formation in the SOSB. Stars mark a series of 27-norcholestanes. At right, values in parentheses represent a measure of relative signal intensity for the C_{26}–C_{30} steranes in acquired MRM chromatograms (though absolute abundances are determined from individual peak areas) and the numbers beneath are the masses (in daltons) of the ion transitions (molecular weight → fragment ion) used in MRM GC-MS in each case. *y* axis, signal intensity; *x* axis, retention time in min (52 to 68 min shown for all traces).

biomarker appearance corresponds well to divergence estimates for the last common ancestor of all living demosponges obtained from molecular clocks[25,26], and indeed can now be used to more robustly calibrate the molecular clock at the base of the animal tree[1].

The use of recalcitrant lipid biomarkers offers a promising approach for tracking the earliest sponge contributions to Precambrian sedimentary rocks because outstanding preservation of soft-body parts, as detected in Doushantuo phosphorites[19,22], is rare in the geological record. Siliceous sponge spicules are metastable and they can be difficult to isolate and identify unambiguously in clastic sediments. Moreover, several orders of Demospongiae completely lack mineral skeletons. On the other hand, the studies of the lipid compositions of Porifera show a remarkable diversity of distinctive structures with abundance patterns aligned to phylogeny[13,27,28]

The demosponge biomarker record for the Huqf Supergroup supports the hypothesis that Metazoa first achieved ecological prominence in shallow marine waters of the Cryogenian[1]. It has been proposed that Neoproterozoic sponges and rangeomorphs feeding on reactive dissolved or particulate marine organic matter[29] may have progressively oxygenated their benthic environments as they moved from shallow water into deeper waters[24]. Consistent with this, our data (Table 1 and Suplementary Table 1) show that, on average, C_{30} steranes comprised 2.7% of total C_{27}–C_{30} extractable steranes in Huqf samples and 63% of the summed C_{30} compounds were 24-isopropylcholestanes, suggesting that demosponges must have made a significant contribution to preserved sedimentary organic matter and, therefore, environmental biomass[24]. In contrast, lack of significant sponge steranes in deepwater shales from the Ediacaran Rodda Bed Formation in the Officer basin, Australia[15], and from the late Cryogenian Aralka Formation (Supplementary Information) suggests that it took longer to colonize deepwater environments. Neoproterozoic sponges would have been at least partly responsible for the ultimate respiration and removal of dissolved organic carbon[24,29], aiding ventilation of the global ocean and shifts in the modes of carbon and sulphur cycling evident from Ediacaran isotopic and geochemical records[3,30].

METHODS SUMMARY

Solvent-rinsed core rock fragments and cuttings were crushed to a fine powder using an alumina ceramic puck mill housed in a SPEX 8510 shatterbox. Rock powders were extracted with a mixture of dichloromethane and methanol (9:1, v/v) using a Dionex Accelerator Solvent Extractor ASE-200 operated under 1,000 p.s.i. at 100 °C. Asphaltenes were precipitated from the resulting organic extracts (bitumens) using *n*-pentane. The maltenes (*n*-pentane solubles) were then fractionated by silica gel adsorption chromatography, eluting successively with hexane, hexane/CH_2Cl_2 (v/v: 4:1) and CH_2Cl_2/CH_3OH (v/v: 3:1) to yield saturated hydrocarbons, aromatic hydrocarbons and resin fractions, respectively.

Continuous-flow hydropyrolysis experiments were conducted on 100–2,000 mg of catalyst-loaded pre-extracted sediments or kerogen concentrates as described previously[7]. Hydropyrolysates were fractionated on silica gel columns, as for rock bitumens.

GC-MS analyses of saturated hydrocarbon fractions were performed on a Micromass AutoSpec Ultima equipped with a HP6890 gas chromatograph and a DB-1MS coated capillary column (60 m × 0.25 mm i.d., 0.25-μm film thickness) using He as carrier gas. Hopane and sterane biomarkers were analysed by MRM GC-MS with a total cycle time of 1.3 s per scan for 26 transitions, including the *m/z* 414 to 217 transition for C_{30} desmethylsteranes. The GC oven was programmed at 60 °C (2 min), heated to 150 °C at 10 °C min^{-1}, further heated to 315 °C at 3 °C min^{-1} and held at final temperature for 24 min.

50 ng of deuterated C_{29} sterane standard [d_4-ααα-24-ethylcholestane (20R)] was typically added to 1 mg saturates to quantify the polycyclic biomarker content. Yields assume equal mass spectral response factors between analytes. Analytical errors for individual hopanes and steranes concentrations are estimated at ±30%. Average uncertainties in hopane and sterane biomarker ratios are ±8% as calculated from multiple analyses of a saturated hydrocarbon fraction from an AGSO standard oil ($n = 30$).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 September; accepted 27 November 2008.

1. Peterson, K. J., Cotton, J. A., Gehling, J. G. & Pisani, D. The Ediacaran emergence of bilaterians: Congruence between the genetic and the geological fossil records. *Phil. Trans. R. Soc. B* **363**, 1435–1443 (2008).
2. Bowring, S. A. *et al.* Geochronologic constraints on the chronostratigraphic framework of the Neoproterozoic Huqf Supergroup, Sultanate of Oman. *Am. J. Sci.* **307**, 1097–1145 (2007).

3. Fike, D. A., Grotzinger, J. P., Pratt, L. M. & Summons, R. E. Oxidation of the Ediacaran ocean. *Nature* **444,** 744–747 (2006).
4. McCarron, G. *The Sedimentology and Chemostratigraphy of the Nafun Group, Huqf Supergroup, Oman.* Thesis, Univ. Oxford (2000).
5. Grantham, P. J., Lijmbach, J., Posthuma, J., Hughes Clarke, M. W. & Willink, R. J. Origin of crude oils in Oman. *J. Petrol. Geol.* **11,** 61–80 (1988).
6. Höld, I. M., Schouten, S., Jellema, J. & Sinninghe Damste, J. S. Origin of free and bound mid-chain methyl alkanes in oils, bitumens and kerogens of the marine, Infracambrian Huqf Formation (Oman). *Org. Geochem.* **30,** 1411–1428 (1999).
7. Love, G. D., Snape, C. E., Carr, A. D. & Houghton, R. C. Release of covalently-bound biomarkers in high yields from kerogen via catalytic hydropyrolysis. *Org. Geochem.* **23,** 981–986 (1995).
8. Murray, I. P., Love, G. D., Snape, C. E. & Bailey, N. J. L. Comparison of covalently-bound aliphatic biomarkers released via hydropyrolysis with their solvent-extractable counterparts for a suite of Kimmeridge clays. *Org. Geochem.* **29,** 1487–1505 (1998).
9. Brocks, J. J. *et al.* Biomarker evidence for green and purple sulphur bacteria in a stratified Palaeoproterozoic sea. *Nature* **437,** 866–870 (2005).
10. Summons, R. E., Jahnke, L. L., Hope, J. M. & Logan, G. A. 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature* **400,** 554–556 (1999).
11. Volkman, J. K. Sterols in microorganisms. *Appl. Microbiol. Biotechnol.* **60,** 495–506 (2003).
12. Moldowan, J. M. *et al.* Sedimentary 24-*n*-propylcholestanes, molecular fossils diagnostic of marine algae. *Science* **247,** 309–312 (1990).
13. Bergquist, P. R., Hofheinz, W. & Oesterhelt, G. Sterol composition and classification of the Demospongiae. *Biochem. Syst. Ecol.* **8,** 423–435 (1980).
14. Silva, C. J., Wunsche, L. & Djerassi, C. Biosynthetic studies of marine lipids 35. The demonstration of de novo sterol biosynthesis in sponges using radiolabelled isoprenoid precursors. *Comp. Biochem. Physiol. B* **99,** 763–773 (1991).
15. McCaffrey, M. A. *et al.* Paleoenvironmental implications of novel C_{30} steranes in Precambrian to Cenozoic age petroleum and bitumen. *Geochim. Cosmochim. Acta* **58,** 529–532 (1994).
16. Kodner, R. B., Summons, R. E., Pearson, A., King, N. & Knoll, A. H. Sterols in a unicellular relative of the metazoans. *Proc. Natl Acad. Sci. USA* **105,** 9897–9902 (2008).
17. Narbonne, G. M. & Gehling, J. G. Life after snowball: The oldest complex Ediacaran fossils. *Geology* **31,** 27–30 (2003).
18. Droser, M. L., Jensen, S. & Gehling, J. G. Trace fossils and substrates of the terminal Proterozoic-Cambrian transition: Implications for the record of early bilaterians and sediment mixing. *Proc. Natl Acad. Sci. USA* **99,** 12572–12576 (2002).
19. Yin, L. *et al.* Doushantuo embryos preserved inside diapause egg cysts. *Nature* **446,** 661–663 (2007).
20. Gehling, J. G. & Rigby, J. K. Long expected sponges from the Neoproterozoic Ediacaran fauna of South Australia. *J. Paleontol.* **70,** 185–195 (1996).
21. Brasier, M., Green, O. & Shields, G. Ediacarian sponge spicule clusters from southwestern Mongolia and the origins of the Cambrian fauna. *Geology* **25,** 303–306 (1997).
22. Li, C. W., Chen, J. Y. & Hua, T. E. Precambrian sponges with cellular structures. *Science* **279,** 879–882 (1998).
23. Xiao, S., Zhang, Y. & Knoll, A. H. Three-dimensional preservation of algae and animal embryos in a Neoproterozoic phosphorite. *Nature* **391,** 553–558 (1998).
24. Sperling, E. A., Peterson, K. J. & Pisani, D. in *The Rise and Fall of the Ediacaran Biota* (eds Vickers-Rich, P. & Komarower, P.) 355–368 (Geological Society Special Publications, 2007).
25. Peterson, K. J. *et al.* Estimating metazoan divergence times with a molecular clock. *Proc. Natl Acad. Sci. USA* **101,** 6536–6541 (2004).
26. Peterson, K. J. & Butterfield, N. J. Origin of the Eumetazoa: Testing ecological predictions of molecular clocks against the Proterozoic fossil record. *Proc. Natl Acad. Sci. USA* **102,** 9547–9552 (2005).
27. Thiel, V. *et al.* A chemical view of the most ancient metazoa – biomarker chemotaxonomy of hexactinellid sponges. *Naturwissenschaften* **89,** 60–66 (2002).
28. Bergquist, P. R., Karuso, P., Cambie, R. C. & Smith, D. J. Sterol composition and classification of the Porifera. *Biochem. Syst. Ecol.* **19,** 17–24 (1991).
29. Rothman, D. H., Hayes, J. M. & Summons, R. E. Dynamics of the Neoproterozoic carbon cycle. *Proc. Natl Acad. Sci. USA* **100,** 8124–8129 (2003).
30. McFadden, K. A. *et al.* Pulsed oxidation and biological evolution in the Ediacaran ocean. *Proc. Natl Acad. Sci. USA* **105,** 3197–3202 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Funding support for this work came from Petroleum Development Oman (PDO), the NASA Exobiology Program, the NSF EAR Program, the Agouron Institute and the NASA Astrobiology Institute. We thank PDO for access to sample materials and Z. Rawahi and P. Taylor, in particular, for their input. C. Colonero, R. Kayser and A. Lewis provided laboratory assistance, including the maintenance of mass spectrometers at MIT.

Author Contributions G.D.L. interpreted the data and wrote the manuscript with input from R.E.S, D.A.F, A.S.B and E.G. G.D.L., E.G., C.S. and A.E.K. acquired the Huqf biomarker data working in the research group of R.E.S. A.S.B. and M.B. screened extant demosponges for their sterol contents. C.E.S. and W.M. made facilities available for HyPy experiments on kerogens and trained C.S. to use the equipment. J.P.G. provided a robust stratigraphic framework for the Huqf Supergroup in the SOSB and with D.A.F. identified key sedimentary rock samples to use in this investigation. S.A.B and D.J.C. measured important U–Pb ages for ash beds and detritral zircons through the stratigraphy to constrain the age range and distribution of our demosponge biomarkers.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to G.D.L. (glove@ucr.edu).

METHODS

The outer surfaces of sediment core and larger cuttings fragments were cleaned sequentially by ultrasonication in distilled water, then methanol, then dichloromethane, and finally *n*-hexane for ~20 s per step before extraction. Cleaned core fragments and cuttings were then crushed to a fine powder using an alumina ceramic puck mill housed in a SPEX 8510 shatterbox. Between samples, the puck mill was cleaned by crushing annealed sand three times for 1-min periods each, followed by washing with the same cleaning solvent sequence described above.

Rock powders were extracted with a mixture of dichloromethane and methanol (9:1, v/v) using a Dionex Accelerator Solvent Extractor ASE-200 operated under 1,000 p.s.i. at 100 °C. Asphaltenes were precipitated out from the resulting organic extracts (bitumens) and from the oils using *n*-pentane. In asphaltene-free fractions (maltenes) derived from bitumens, elemental sulphur was removed with activated and solvent-washed copper pellets. The maltenes were then fractionated by silica gel column chromatography eluting successively with hexane, hexane/CH_2Cl_2 (v/v: 4:1) and CH_2Cl_2/CH_3OH (v/v: 3:1) to yield saturated hydrocarbons, aromatic hydrocarbons and polars/resins (N, S, O compounds), respectively.

Continuous-flow HyPy experiments were performed on 100–2,000 mg of catalyst-loaded pre-extracted sediments or kerogen concentrates as described previously[7]. The isolation of kerogen concentrates was conducted on solvent-extracted rock residues by standard hydrofluoric acid/hydrochloric acid (HF/HCl) extraction procedures. Further treatment of the isolated kerogens involved extraction with dichloromethane by ultrasonication (×3). Extracted sediments and kerogens were initially impregnated with an aqueous methanol solution of ammonium dioxydithiomolybdate, $(NH_4)_2MoO_2S_2$, to give a give a nominal loading of 2 wt% molybdenum. Ammonium dioxydithiomolybdate reductively decomposes *in situ* under HyPy conditions above 250 °C to form a catalytically active molybdenum sulphide (MoS_2) phase. The catalyst-loaded samples were heated in a stainless steel (316 grade) reactor tube from ambient temperature to 260 °C at 300 °C min^{-1} then to 520 °C at 8 °C min^{-1}. A hydrogen sweep gas flow of 6 dm^3 min^{-1}, measured at ambient temperature and pressure, through the reactor bed ensured that the residence times of volatiles generated was the order of only a few seconds. Products were collected in a silica gel trap cooled with dry ice and the adsorbed pyrolysates were separated into saturates, aromatics and polars using silica gel column chromatography as for rock bitumens. Solvent-extracted, activated copper turnings were added to concentrated solutions of saturated hydrocarbon fractions to remove all traces of elemental sulphur, which is formed from disproportionation of the catalyst during HyPy.

To reduce the levels of background contamination in HyPy, a cleaning run was performed before each sample run whereby the apparatus was heated to 520 °C using a rapid heating rate (300 °C min^{-1}) under high-hydrogen-pressure conditions. Experimental blanks, using annealed silica gel in the reactor tube instead of a kerogen sample, were regularly performed and the products monitored and quantified to ensure that trace organic contamination levels were acceptably low.

For a sub-set of the rock extracts, branched and cyclic saturated hydrocarbons were separated from straight-chain alkanes by treating the saturated hydrocarbon fraction with silicalite molecular sieve. Approximately 5–10 mg of saturated hydrocarbons, dissolved in a minimum volume of *n*-pentane, was placed on a 3 cm bed of activated, crushed silicalite lightly packed into a Pasteur pipette. The silicalite non-adduct (SNA) containing branched and cyclic alkanes was washed through using pentane (4 ml).

A deuterated C_{29} sterane standard (d4-ααα-24-ethylcholestane (20R), Chiron Laboratories AS) was added to branched/cyclic alkane or total saturate fractions before GC-MS to quantify biomarker peaks, with typically 50 ng internal standard added to a 1 mg aliquot of saturates. In MRM analyses, this standard compound was detected using the *m/z* 404 to 221 transition.

GC-MS analyses on saturated hydrocarbon fractions were carried out on a Micromass AutoSpec Ultima equipped with a HP6890 gas chromatograph (Hewlett Packard) and a DB-1MS coated capillary column (60 m × 0.25 mm i.d., 0.25-μm film thickness) using He as carrier gas. The MS source was operated at 250 °C in EI mode at 70-eV ionization energy and with 8,000-V acceleration voltage. Samples were injected in pulsed splitless mode into a Gerstel PTV injector at a constant temperature of 300 °C. For full-scan and selected ion recording (SIR) experiments, the GC oven was programmed at 60 °C (2 min), heated to 315 °C at 4 °C min^{-1}, with a final hold time of 35 min. The AutoSpec full-scan duration was 0.8 s plus 0.2 s interscan delay over a mass range of 50 to 600 Da. Hopane and sterane biomarkers were analysed by MRM GC-MS with a total cycle time of 1.3 s per scan for 26 parent-fragment transitions, including the *m/z* 414 to 217 transition for C_{30} desmethylsteranes. For MRM, the GC oven was programmed at 60 °C (2 min), heated to 150 °C at 10 °C min^{-1}, further heated to 315 °C at 3 °C min^{-1} and held at the final temperature for 24 min.

Peak identifications of 24-isopropylcholestanes were confirmed by comparison of retention times with an AGSO oil saturated hydrocarbon standard and with Neoproterozoic oils from Siberia[15] shown previously to contain significant quantities of 24-isopropylcholestanes. Polycyclic biomarkers were quantified assuming equal mass spectral response factors between analytes and the d_4-C_{29}-ααα-ethylcholestane (20R) internal standard. Analytical errors for absolute yields of individual hopanes and steranes are estimated at ±30%. Average uncertainties in hopane and sterane biomarker ratios are ±8% as calculated from multiple analyses of a saturated hydrocarbon fraction prepared from an AGSO standard oil ($n = 30$ MRM analyses).

Vol 457|5 February 2009|doi:10.1038/nature07537

LETTERS

A human natural killer cell subset provides an innate source of IL-22 for mucosal immunity

Marina Cella[1]*, Anja Fuchs[1]*, William Vermi[2], Fabio Facchetti[2], Karel Otero[1], Jochen K. M. Lennerz[1], Jason M. Doherty[1], Jason C. Mills[1] & Marco Colonna[1]

Natural killer (NK) cells are classically viewed as lymphocytes that provide innate surveillance against virally infected cells and tumour cells through the release of cytolytic mediators and interferon (IFN)-γ. In humans, blood CD56^{dim} NK cells specialize in the lysis of cell targets[1]. In the lymph nodes, CD56^{bright} NK cells secrete IFN-γ cooperating with dendritic cells and T cells in the generation of adaptive responses[1,2]. Here we report the characterization of a human NK cell subset located in mucosa-associated lymphoid tissues, such as tonsils and Peyer's patches, which is hard-wired to secrete interleukin (IL)-22, IL-26 and leukaemia inhibitory factor. These NK cells, which we refer to as NK-22 cells, are triggered by acute exposure to IL-23. *In vitro*, NK-22-secreted cytokines stimulate epithelial cells to secrete IL-10, proliferate and express a variety of mitogenic and anti-apoptotic molecules. NK-22 cells are also found in mouse mucosa-associated lymphoid tissues and appear in the small intestine lamina propria during bacterial infection, suggesting that NK-22 cells provide an innate source of IL-22 that may help constrain inflammation and protect mucosal sites.

Human NK cells have been dissected into CD56^{dim} and CD56^{bright} subsets possessing either lytic or IFN-γ secretory functions[1]. Recently, a subset of tonsil NK cells was shown to express the receptor NKp44 (ref. 2), which is not present on blood NK cells unless they are activated *in vitro* with IL-2 or IL-15 (ref. 3). We noticed that NKp44$^+$ NK cells are present in tonsils, but not in lymph nodes (Supplementary Fig. 1). Immunohistochemical and immunofluorescent analyses showed that NKp44$^+$ NK cells are predominantly located in the mucosa surrounding the lymphoid follicles (Fig. 1a–c), with only a small number in the interfollicular area (data not shown). Because tonsils are associated with the oral mucosa, we conceived that NKp44 identifies a subset of NK cells preferentially situated in mucosa-associated lymphoid tissues (MALT). Indeed, we also found NKp44$^+$ cells in the Peyer's patches of the ileum and the appendix (Fig. 1d, e).

Gene expression profiles of tonsil NKp44$^+$ and NKp44$^-$ NK cells (Supplementary Fig. 2 and Supplementary Table 1) showed that NKp44$^+$ NK cells preferentially express the chemokine receptor CCR6 and its ligand CCL20 (Supplementary Table 1), which are known to direct the mucosal migration of memory CD4$^+$ T cells, B cells, dendritic cells and T_H17 T cells[4,5]. Flow cytometry and migration assays confirmed that a large fraction of NKp44$^+$ NK cells express functional CCR6 (Fig. 2a, b). Moreover, NKp44$^+$ NK cells secreted more CCL20 than NKp44$^-$ NK cells (Fig. 2c). Thus, NKp44$^+$ NK cells can contribute to CCL20 production in the mucosa, promoting their own accumulation and attracting other immune cells. The finding that the CCR6–CCL20 axis has an important role in driving the localization of NKp44$^+$ NK cells was corroborated by the analysis of human dermatitis with proliferation and retention of Langerhans cell in the skin[6]. Langerhans cell accumulation caused abnormal production of CCL20, resulting in extensive infiltration of NKp44$^+$ NK cells (Fig. 2d–f). Consistent with their localization in MALT, NKp44$^+$ NK cells also expressed several proteins known to promote lymphocyte adhesion to epithelial cells, including CD96 (ref. 7) and CD103 (ref. 8; Supplementary Table 1). Flow cytometric analysis confirmed higher levels of CD96 on all NKp44$^+$ NK cells than on NKp44$^-$ NK cells (Fig. 2g) as well as expression of CD103 on a considerable percentage of NKp44$^+$ NK cells (Fig. 2h). Moreover, NKp44$^+$ NK cells adhered to epithelial cells better than to NKp44$^-$ NK cells (Fig. 2i, j).

Next we evaluated the function of NKp44$^+$ NK cells. Previous studies have demonstrated that tonsil NK cells have limited cytotoxic capacity[2]. Indeed, no intracellular perforin and little intracellular granzyme B were detected in NKp44$^+$ NK cells (Supplementary Fig. 3). Moreover, intracellular IFN-γ was barely detectable in NKp44$^+$ NK cells (Supplementary Fig. 4). Notably, gene chip analysis of NKp44$^+$ NK cells revealed production of IL-22, which is involved in mucosal defence[9–13]. IL-26 and leukaemia inhibitory factor (LIF), which activate epithelial cells[14], were also increased (Supplementary Table 1). To corroborate these data, we measured IL-22 secretion by tonsil NK cells stimulated with PMA and ionomycin and assessed the

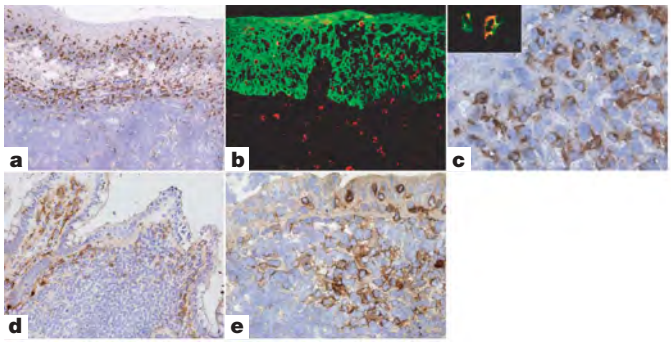


Figure 1 | NKp44$^+$ NK cells are prominently found within MALT. **a**, Immunohistochemical analysis of tonsil sections with anti-NKp44 shows NKp44$^+$ NK cells residing within the tonsil epithelium and the lamina propria. **b**, Double immunofluorescence confirms that NKp44$^+$ cells (red) are either in the lamina propria, or within the surface epithelium in close contact with cytokeratin-5$^+$ (green) epithelial cells. **c**, At high-power magnification, NKp44$^+$ NK cells show a round to oval morphology and co-express CD56 (inset). **d**, **e**, Numerous NKp44$^+$ NK cells are present in the dome region of the ileum Peyer's patches (**d**) and in the luminal epithelium, dome and interfollicular regions of the appendix (**e**). Original magnifications were ×100 (**a**), ×200 (**b**, **d** and inset in **c**), ×400 (**e**) and ×600 (**c**).

[1]Department of Pathology and Immunology, Washington University School of Medicine, St Louis, Missouri 63110, USA. [2]Department of Pathology I, Spedali Civili, University of Brescia, 25123 Brescia, Italy.
*These authors contributed equally to this work.

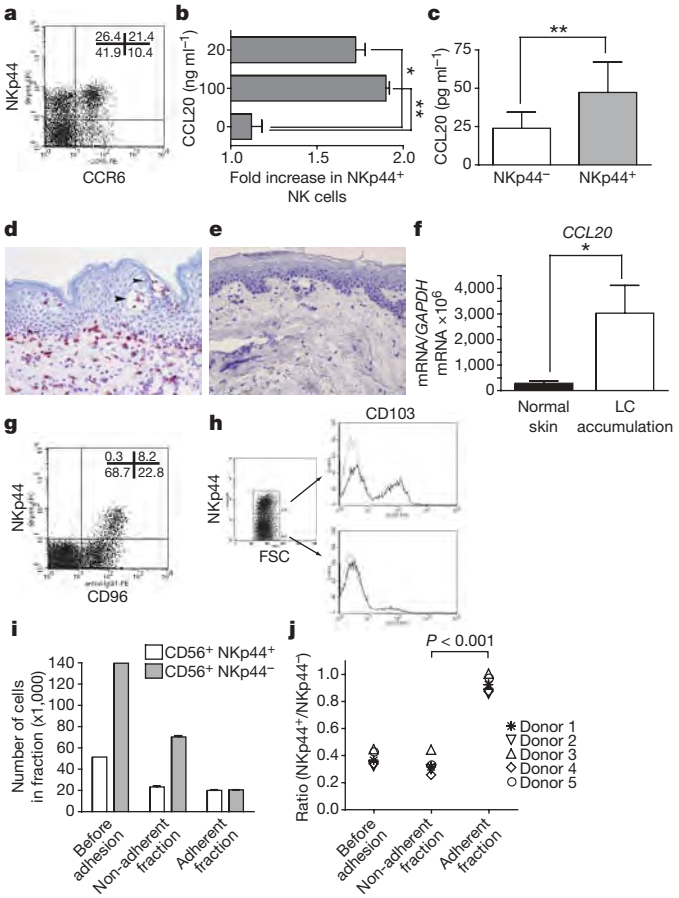


Figure 2 | A subset of tonsil NKp44$^+$ NK cells express CCR6, respond to CCL20 *in vitro* and *in vivo*, produce CCL20 and adhere to epithelial cells. **a**, Approximately 50% of NKp44$^+$ NK cells express CCR6. **b**, Tonsil NK cells migrate in response to CCL20. Bars indicate the ratio between the percentage of input NKp44$^+$ NK cells versus the percentage of migrated NKp44$^+$ NK cells. **c**, PMA-and-ionomycin-treated NKp44$^+$ NK cells produce significantly more CCL20 than NKp44$^-$ NK cells. **d**, **e**, A case of dermatitis with pathological Langerhans cell accumulation shows extensive skin infiltration of NKp44$^+$ NK cells as compared to normal skin. Arrowheads indicate a pseudo-Pautrier abscess indicative of Langerhans cell accumulation[6]. Original magnifications were ×200. **f**, Real-time PCR shows that skin with Langerhans cell (LC) accumulation produces more *CCL20* than normal skin. **g**, NKp44$^+$ NK cells express higher levels of CD96 than NKp44$^-$ NK cells. **h**, A discrete subset of NKp44$^+$ NK cells expresses CD103. FSC, forward scatter. **i**, NKp44$^+$ NK cells firmly adhere to epithelial cells. The ratio of NKp44$^+$ to NKp44$^-$ cells within non-adherent and adherent CD3$^-$CD56$^+$ cells was determined by flow cytometry. **j**, The adhesiveness of NKp44$^+$ NK cells to epithelial cells is consistent in different donors. $^*P < 0.05$; $^{**}P < 0.01$; error bars, s.d.; $n = 3$ (**b**), 8 (**c**) and 6 (**f**).

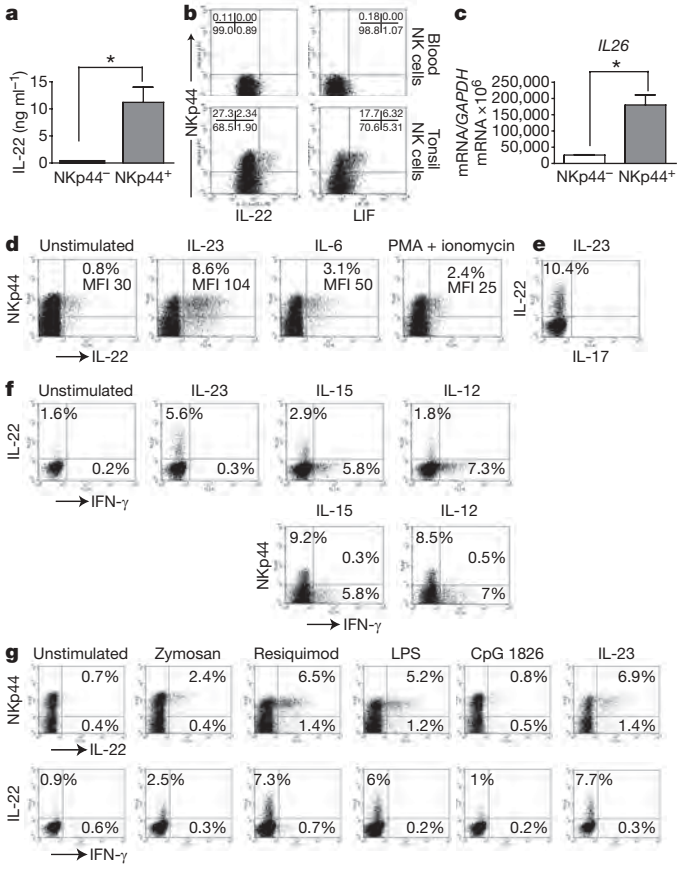


Figure 3 | Tonsil NKp44$^+$ NK cells produce IL-22, IL-26 and LIF. IL-23 and TLR-activated monocytes trigger IL-22 secretion. **a**, IL-22 concentration in supernatants of NKp44$^+$ and NKp44$^-$ tonsil NK cells stimulated with PMA and ionomycin. **b**, Intracellular IL-22 and LIF content of tonsil NKp44$^+$ cells and blood NK cells stimulated with PMA and ionomycin. **c**, NKp44$^+$ NK cells express higher levels of *IL26* mRNA than NKp44$^-$ NK cells. **d**, **e**, IL-23 triggers the secretion of IL-22 (**d**, **e**) but not IL-17 (**e**) by NK-22 cells. MFI, mean fluorescence intensity. **f**, NKp44$^+$ NK cells produce IL-22 in response to IL-23 (10 ng ml^{-1}) and, to some extent, to IL-15 (10 ng ml^{-1}). NKp44$^-$ NK cells release IFN-γ in response to IL-12 or IL-15. **g**, Monocytes stimulated with resiquimod and LPS induce IL-22 secretion in NK-22 cells. $^*P < 0.05$; error bars, s.d.; $n = 5$ (**a**) and 2 (**c**).

intracellular content of IL-22 and LIF in activated tonsil NK cell subsets. Both assays confirmed augmented IL-22 and LIF secretion by NKp44$^+$ NK cells (Fig. 3a, b). Preferential *IL-26* expression by NKp44$^+$ NK cells was validated by real-time polymerase chain reaction (PCR) analysis (Fig. 3c). Although IL-22 and IL-26 are part of the T_H17 CD4$^+$ T cell cytokine profile[15–19], NKp44$^+$ NK cells did not produce IL-17 (data not shown). Thus, we will refer to NKp44$^+$ NK cells that secrete IL-22 as NK-22 cells.

Although PMA and ionomycin stimulation exposed the unique NK-22 cytokine profile, the physiological stimuli that activate NK-22 cells were not clear. Engagement of NKp44 or other receptors with specific antibodies had no effect (data not shown). However, stimulation of tonsil NKp44$^+$ NK cells with a variety of inflammatory cytokines revealed that NK-22 cells are selectively and acutely responsive to IL-23 (Fig. 3d and Supplementary Fig. 5a). Similar results were observed for Peyer's patch NK cells (Supplementary Fig. 5b). Of all NKp44$^+$ NK cells, only those expressing CCR6 responded to IL-23 (Supplementary Fig. 5c), demonstrating that NK-22 cells are a discrete subset within NKp44$^+$ NK cells. IL-23 did not elicit IL-17 (Fig. 3e). NK-22 cells were also slightly responsive to IL-6 and IL-15 (Fig. 3d, f). IL-12, which induces IFN-γ in blood NK cells, induced neither IL-22 nor IFN-γ in NK-22 cells (Fig. 3f).

In contrast, tonsil NKp44$^-$ NK cells did not respond to IL-23, but did react to IL-12, secreting IFN-γ rather than IL-22 (Fig. 3f). In addition, IL-15 and IL-12 induced IFN-γ in NKp44$^-$ NK cells almost equally well (Fig. 3f). Neither IL-23 nor IL-15 induced IL-22 by peripheral blood NK cells (data not shown). Together, these results indicate that MALT NK-22 cells are hard-wired to secrete IL-22, particularly in response to IL-23. In contrast, NKp44$^-$ NK cells are programmed to secrete IFN-γ after stimulation with IL-12 or IL-15. Thus, NKp44$^-$ NK cells probably correspond to the tonsil NK cells recently shown to inhibit Epstein–Barr virus-induced B cell transformation through IFN-γ secretion[20].

Monocytes, dendritic cells and macrophages produce IL-23 in response to microbial stimuli[21] and hence may provide a principal intramucosal source of stimuli for NK-22 cell activation. To test this hypothesis, tonsil NK cells were co-cultured with either unstimulated or activated human allogeneic monocytes. As predicted, activated

monocytes induced IL-22 production by NK-22 cells (Fig. 3g). Culture supernatants from activated monocytes contained IL-23 and induced IL-22 secretion just like monocytes (Supplementary Fig. 6).

To assess the effect of NK-22 cell-secreted cytokines on epithelial cell function, we measured proliferation and cytokine secretion of colon epithelial cells after stimulation with culture supernatant collected from activated NK-22 cells. NK-22 cell-conditioned medium induced the proliferation of epithelial cells (Fig. 4a) and the secretion of the anti-inflammatory cytokine IL-10 (Fig. 4b). Importantly, NK-22 cell-conditioned medium was more effective than recombinant IL-22 alone, consistent with a combined action of IL-22, IL-26 and possibly yet unknown factors. IL-22, IL-26 and LIF have been shown to activate signal transducer and activator of transcription 3 (STAT3)[10,22], a known inducer of cell survival and proliferation, as well as STAT1. Robust STAT3 and STAT1 phosphorylation was observed in epithelial cells after stimulation with NK-22 cell-conditioned medium (Fig. 4c). To further define downstream signalling mediators induced by NK-22-derived cytokines, we obtained gene expression profiles of colon epithelial cells treated with recombinant IL-22 (Supplementary Table 2). IL-22-induced genes were subsequently verified in epithelial cells stimulated with NK-22 cell-conditioned medium by immunoblot and real-time PCR analyses. Notably, NK-22 cell-conditioned medium strongly induced the proto-oncogenes *Bcl-3* and *Bcl-6* (Supplementary Fig. 7). Bcl-3 interaction with NF-κB promotes the transcription of proliferation genes in response to growth signals, whereas Bcl-6 may favour proliferation by preventing terminal differentiation[23]. Furthermore, we detected transcriptional activation of genes involved in cell growth (*PBEF1*, also known as *NAMPT*), cell cycle progression (*MYC*) and protection from apoptotic stimuli (kinase *SGK1*, serine protease inhibitors *SERPINA1* and *SERPINA3*, and *GADD45G*) (Supplementary Fig. 7). Thus, NK-22 cytokines stimulate epithelial cell proliferation, secretion of IL-10 and activate a variety of mitogenic and anti-apoptotic pathways.

To identify NK-22 cells in mouse MALT, we examined NK cells in Peyer's patches of C57BL/6 mice and found two subsets: NKp46$^+$ NK1.1$^+$ (40–50% of total NKp46$^+$ cells) and NKp46$^+$ NK1.1$^-$ (Supplementary Fig. 8). Acute *in vitro* stimulation with IL-23 elicited production of IL-22 in both subsets (Fig. 5a, b). No NK-22 cells were detectable in the lamina propria or within the intestinal epithelium under homeostatic conditions (Supplementary Fig. 9a). Because IL-22 has been shown to be essential for the early host defence against *Citrobacter rodentium*[11], we infected mice with *C. rodentium* and observed the appearance of NK-22 cells in the lamina propria 6 days after oral inoculation (Fig. 5c and Supplementary Fig. 9b). In recombination activating gene-2 (*Rag2*)-deficient mice, which lack T-cell sources of IL-22, NK-22 cells were detected in the intestinal epithelium before and after infection with *C. rodentium* (Supplementary Fig. 10a) and spontaneously secreted IL-22 *ex vivo* after infection. Moreover, depletion of *Rag2*$^{-/-}$ mice with NK1.1 at the early stages of infection resulted in accelerated mortality of infected mice (Supplementary Fig. 10b), indicating that NK-22 may provide an innate source of IL-22 that contributes to mucosal immunity.

Finally, we investigated the developmental pathway of NK-22 cells. Several similarities between NK-22 and T_H17 cells suggest that there may be similar differentiation requirements. This was further supported

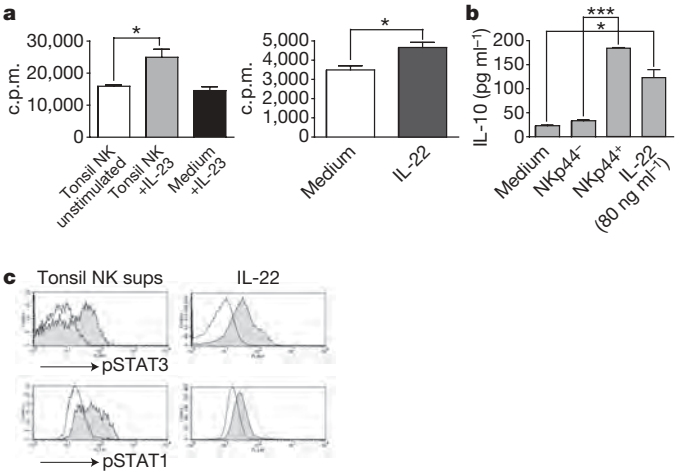


Figure 4 | NK-22 cell-secreted cytokines stimulate epithelial cells to proliferate, release IL-10 and activate STAT1 and STAT3. **a**, Colo205 cell proliferation in response to supernatants derived from resting or IL-23-stimulated tonsil NK cells (left panel) or IL-22 (right panel). **b**, IL-10 secretion by Colo205 cells stimulated with supernatants from activated NKp44$^+$ or NKp44$^-$ tonsil NK cells or IL-22. **c**, STAT3 and STAT1 phosphorylation in Colo205 cells stimulated with supernatants of IL-23-activated NKp44$^+$ NK cells (left panels, grey histograms), NKp44$^-$ NK cells (left panels, empty histograms), IL-22 (right panels, grey histograms) or control medium (right panels, empty histograms). As an extra control for STAT1 phosphorylation, cells were stimulated with IFN-γ (dotted line). Sups, supernatants; $*P < 0.05$; $***P < 0.0001$; error bars, s.d.; $n = 6$ (**a**) and 3 (**b**).

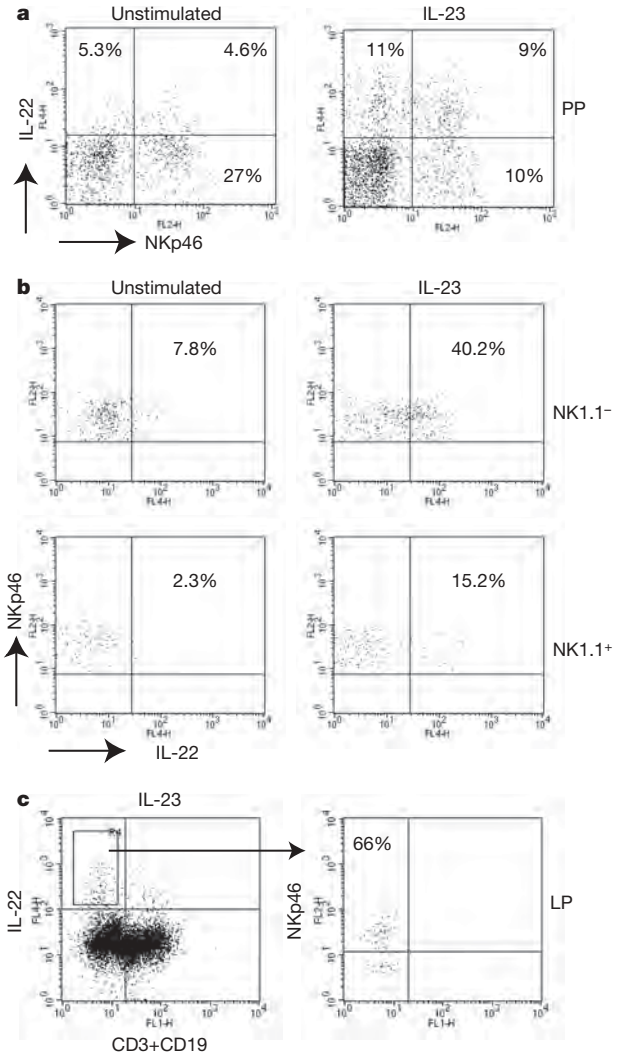


Figure 5 | Identification of mouse NK-22 cells. **a**, *In vitro* stimulation of Peyer's patches (PP) cells with IL-23 induces production of IL-22 by NKp46$^+$ NK cells. A gate was applied to exclude CD3$^+$CD19$^+$ cells. **b**, NKp46$^+$ IL-22-producing cells are in part NK1.1$^-$ and, to a lower degree, NK1.1$^+$. **c**, NK-22 cells appear in the small intestine lamina propria of *C. rodentium*-infected mice. A large proportion (55–80% in different experiments) of cells producing IL-22 in the lamina propria (LP) express NKp46.

by NK-22 expression of the aryl hydrocarbon receptor (AHR), ROR-α, ROR-γ and IRF4 (Supplementary Fig. 11 and Supplementary Table 1), which are all transcription factors recognized as being required for T_H17 differentiation and/or IL-22 secretion[24–26]. However, the culture of peripheral blood or cord blood NK cells under various T_H17 polarizing conditions[4,5,15,27–29] with or without various AHR ligands[24,25] did not yield any IL-22-producing NK cells. Thus, NK-22 cells may develop from local progenitors, perhaps lymphoid tissue inducer cells, which also express ROR-γ[30], under the influence of microenvironmental factors such as intestinal microbiota.

We propose that NK-22 cells are an innate cell type that contributes to mucosal homeostasis. Although NK-22 and T_H17 cells share several features, NK-22 cells are unique in several respects. First, because they do not secrete IL-17, NK-22 cells are unlikely to have pro-inflammatory activity. Second, whereas T_H17 cells are triggered through the T-cell receptor and require IL-23 for expansion[17–19], NK-22 cells are directly and acutely stimulated by IL-23. Third, blood NK cells cultured in T_H17 polarizing cytokines and AHR agonists do not acquire the ability to secrete IL-22, perhaps reflecting a distinct differentiation pathway. We suggest that after microbial challenge of mucosal barriers and the subsequent release of IL-23 by antigen-presenting cells, NK-22 cells provide a source of IL-22 and other cytokines that may help constrain inflammation and protect mucosal sites.

METHODS SUMMARY

Isolation of tonsil NK cells. Tonsils were mechanically disrupted and NK cells were pre-enriched by magnetic purification with CD56 microbeads (Miltenyi Biotec). For many experiments cells were further stained with a combination of anti-CD56, anti-CD3, anti-CD19 (Pharmingen) and anti-NKp44 antibodies and sorted either on a Moflo cell sorter (Cytomation) or a FACSVantage sorter (BD Biosciences).

Reagents. Human and mouse IL-22 ELISA kits were obtained from Antigenix America. The CCL20 ELISA kit, recombinant CCL20, IL-23 ELISA kit, anti-human IL-22 antibody for intracellular staining and anti-mouse NKp46 were obtained from R&D. IL-22 and IL-1β were purchased from Peprotec. IL-22 was used at 80 ng ml^{-1} for stimulation of epithelial cells. IL-23 was used at 40 ng ml^{-1} for stimulation of NK cells, unless otherwise stated. IL-6, IL-15, IL-26, mouse and human IL-23 were purchased from R&D. Anti-IL-17A antibody was obtained from eBioscience. IL-12p70 and antibodies against IFN-γ, TNF-α, GM-CSF, LIF, perforin, granzyme, CCR6, CD103, pSTAT3, pSTAT1, NK1.1, mouse CD3 and mouse CD19 were purchased from Pharmingen. Antibodies against BCL-3, BCL-6, MCL-1 and actin for immunoblotting were obtained from Santa Cruz Biotechnology. IL-10 was quantified in cell culture supernatants with the CBA human inflammation kit (Pharmingen). Anti-NKp44 and anti-CD96 antibodies were produced and conjugated in our laboratory.

Received 24 September; accepted 15 October 2008.
Published online 2 November 2008.

1. Cooper, M. A., Fehniger, T. A. & Caligiuri, M. A. The biology of human natural killer-cell subsets. *Trends Immunol.* **22,** 633–640 (2001).
2. Ferlazzo, G. & Munz, C. NK cell compartments and their activation by dendritic cells. *J. Immunol.* **172,** 1333–1339 (2004).
3. Vitale, M. *et al.* NKp44, a novel triggering surface molecule specifically expressed by activated natural killer cells, is involved in non-major histocompatibility complex-restricted tumor cell lysis. *J. Exp. Med.* **187,** 2065–2072 (1998).
4. Acosta-Rodriguez, E. V. *et al.* Surface phenotype and antigenic specificity of human interleukin 17-producing T helper memory cells. *Nature Immunol.* **8,** 639–646 (2007).
5. Manel, N., Unutmaz, D. & Littman, D. R. The differentiation of human T_H-17 cells requires transforming growth factor-β induction of the nuclear receptor RORγt. *Nature Immunol.* **9,** 641–649 (2008).
6. Candiago, E., Marocolo, D., Manganoni, M. A., Leali, C. & Facchetti, F. Nonlymphoid intraepidermal mononuclear cell collections (pseudo-Pautrier abscesses): a morphologic and immunophenotypical characterization. *Am. J. Dermatopathol.* **22,** 1–6 (2000).
7. Fuchs, A., Cella, M., Giurisato, E., Shaw, A. S. & Colonna, M. Cutting edge: CD96 (tactile) promotes NK cell-target cell adhesion by interacting with the poliovirus receptor (CD155). *J. Immunol.* **172,** 3994–3998 (2004).
8. Cepek, K. L. *et al.* Adhesion between epithelial cells and T lymphocytes mediated by E-cadherin and the $\alpha^E\beta_7$ integrin. *Nature* **372,** 190–193 (1994).
9. Aujla, S. J. *et al.* IL-22 mediates mucosal host defense against Gram-negative bacterial pneumonia. *Nature Med.* **14,** 275–281 (2008).
10. Zheng, Y. *et al.* Interleukin-22, a T_H17 cytokine, mediates IL-23-induced dermal inflammation and acanthosis. *Nature* **445,** 648–651 (2007).
11. Zheng, Y. *et al.* Interleukin-22 mediates early host defense against attaching and effacing bacterial pathogens. *Nature Med.* **14,** 282–289 (2008).
12. Sugimoto, K. *et al.* IL-22 ameliorates intestinal inflammation in a mouse model of ulcerative colitis. *J. Clin. Invest.* **118,** 534–544 (2008).
13. Zenewicz, L. A. *et al.* Interleukin-22 but not interleukin-17 provides protection to hepatocytes during acute liver inflammation. *Immunity* **27,** 647–659 (2007).
14. Hor, S. *et al.* The T-cell lymphokine interleukin-26 targets epithelial cells through the interleukin-20 receptor 1 and interleukin-10 receptor 2 chains. *J. Biol. Chem.* **279,** 33343–33351 (2004).
15. Wilson, N. J. *et al.* Development, cytokine profile and function of human interleukin 17-producing helper T cells. *Nature Immunol.* **8,** 950–957 (2007).
16. Liang, S. C. *et al.* Interleukin (IL)-22 and IL-17 are coexpressed by Th17 cells and cooperatively enhance expression of antimicrobial peptides. *J. Exp. Med.* **203,** 2271–2279 (2006).
17. Weaver, C. T., Hatton, R. D., Mangan, P. R. & Harrington, L. E. IL-17 family cytokines and the expanding diversity of effector T cell lineages. *Annu. Rev. Immunol.* **25,** 821–852 (2007).
18. Stockinger, B., Veldhoen, M. & Martin, B. Th17 T cells: linking innate and adaptive immunity. *Semin. Immunol.* **19,** 353–361 (2007).
19. Ouyang, W., Kolls, J. K. & Zheng, Y. The biological functions of T helper 17 cell effector cytokines in inflammation. *Immunity* **28,** 454–467 (2008).
20. Strowig, T. *et al.* Tonsilar NK cells restrict B cell transformation by the Epstein-Barr virus via IFN-γ. *PLoS Pathog.* **4,** e27 (2008).
21. Acosta-Rodriguez, E. V., Napolitani, G., Lanzavecchia, A. & Sallusto, F. Interleukins 1β and 6 but not transforming growth factor-β are essential for the differentiation of interleukin 17-producing human T helper cells. *Nature Immunol.* **8,** 942–949 (2007).
22. Nagalakshmi, M. L., Rascle, A., Zurawski, S., Menon, S. & de Waal Malefyt, R. Interleukin-22 activates STAT3 and induces IL-10 by colon epithelial cells. *Int. Immunopharmacol.* **4,** 679–691 (2004).
23. Kusam, S. & Dent, A. Common mechanisms for the regulation of B cell differentiation and transformation by the transcriptional repressor protein BCL-6. *Immunol. Res.* **37,** 177–186 (2007).
24. Veldhoen, M. *et al.* The aryl hydrocarbon receptor links T_H17-cell-mediated autoimmunity to environmental toxins. *Nature* **453,** 106–109 (2008).
25. Quintana, F. J. *et al.* Control of T_{reg} and T_H17 cell differentiation by the aryl hydrocarbon receptor. *Nature* **453,** 65–71 (2008).
26. Brustle, A. *et al.* The development of inflammatory T_H-17 cells requires interferon-regulatory factor 4. *Nature Immunol.* **8,** 958–966 (2007).
27. Volpe, E. *et al.* A critical function for transforming growth factor-β, interleukin 23 and proinflammatory cytokines in driving and modulating human T_H-17 responses. *Nature Immunol.* **9,** 650–657 (2008).
28. Zhou, L. *et al.* TGF-β-induced Foxp3 inhibits T_H17 cell differentiation by antagonizing RORγt function. *Nature* **453,** 236–240 (2008).
29. Yang, L. *et al.* IL-21 and TGF-β are required for differentiation of human T_H17 cells. *Nature* **454,** 350–352 (2008).
30. Mebius, R. E. Organogenesis of lymphoid tissues. *Nature Rev. Immunol.* **3,** 292–303 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Hughes, B. Eades and S. Schloehmann for cell sorting; R. Clary and the nursing staff at the Children's Hospital, Washington University School of Medicine, for providing tonsil specimens; S. Lonardi for assistance in immunohistochemistry; J. Pfeifer for providing gut specimens; A. Rapaport and S. McCartney for help in gene chip analysis; and S. Gilfillan for critically reading the manuscript. This work was supported by National Institutes of Health (NIH) grants R01AI056139-05 and R21AI067748-02 (to M.Co.) and National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) grant R01 DK079798 (to J.C.M.). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIH and NIDDK.

Author contributions M.Ce., A.F., K.O. and J.M.D. performed the experiments. W.V., F.F. and J.K.M.L. performed immunohistochemical analyses. J.C.M. designed experiments. M.Co. wrote the manuscript and directed the research.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.Co. (mcolonna@pathology.wustl.edu).

LETTERS

Signalling through RHEB-1 mediates intermittent fasting-induced longevity in *C. elegans*

Sakiko Honjoh[1], Takuya Yamamoto[1], Masaharu Uno[1] & Eisuke Nishida[1]

Dietary restriction is the most effective and reproducible intervention to extend lifespan in divergent species[1]. In mammals, two regimens of dietary restriction, intermittent fasting (IF) and chronic caloric restriction, have proven to extend lifespan and reduce the incidence of age-related disorders[2]. An important characteristic of IF is that it can increase lifespan even when there is little or no overall decrease in calorie intake[2]. The molecular mechanisms underlying IF-induced longevity, however, remain largely unknown. Here we establish an IF regimen that effectively extends the lifespan of *Caenorhabditis elegans*, and show that the low molecular weight GTPase RHEB-1 has a dual role in lifespan regulation; RHEB-1 is required for the IF-induced longevity, whereas inhibition of RHEB-1 mimics the caloric-restriction effects. RHEB-1 exerts its effects in part by the insulin/insulin growth factor (IGF)-like signalling effector DAF-16 in IF. Our analyses demonstrate that most fasting-induced upregulated genes require RHEB-1 function for their induction, and that RHEB-1 and TOR signalling are required for the fasting-induced downregulation of an insulin-like peptide, INS-7. These findings identify the essential role of signalling by RHEB-1 in IF-induced longevity and gene expression changes, and suggest a molecular link between the IF-induced longevity and the insulin/IGF-like signalling pathway.

In an IF regimen, which has not been established in invertebrate model organisms, food is provided *ad libitum* to both control and experimental groups, but the experimental group is subjected to periods of fasting. We tested two IF regimens, an alternate-day fasting and an every 2 days fasting in *C. elegans*—an organism that has been shown to be an excellent model system for ageing research—and found that they increased lifespan by 40.4% and 56.6%, respectively (Fig. 1a, b and Supplementary Table 1). Therefore, we used fasting every 2 days as the IF regimen. This IF regimen increased resistance to heat and oxidative stress (Fig. 1c), and markedly delayed the age-related physiological decline. As animals age, the locomotion activity and muscle integrity decrease, showing impairment of cellular functions. IF markedly suppressed the age-dependent decline in these activities (Fig. 1d and Supplementary Fig. 1), suggesting that ageing is delayed in *C. elegans* by the IF regimen established here.

To examine the relationship between IF and caloric restriction, we chose the solid dietary restriction method (dilution of *E. coli* on agar plates) as a caloric-restriction assay to perform IF and caloric restriction in the isogenic backgrounds. Consistent with previous reports, chronic restriction of food intake extended lifespan significantly; the mean lifespan in caloric restriction (5.0×10^8 bacteria ml^{-1}) was increased by 13.2% ($P < 1.1 \times 10^{-5}$, *t*-test) compared to that in *ad libitum* (5.0×10^{10} bacteria ml^{-1}; Fig. 1e). We introduced the IF regimen to these food-restricted animals. Our results showed that the effects of IF and caloric restriction are overlapping. The extent of lifespan extension by IF in *ad libitum* was significantly larger than that in caloric restriction (66.5% versus 51.3%). Furthermore, the mean lifespans of worms subjected to IF under both conditions are not statistically different (41.9 and 43.1 days, respectively; $P = 0.38$, *t*-test). Similarly, IF extended the lifespan of the *clk-1* mutant *clk-1(e2519)*, which is also reported to be long-lived owing to a common caloric-restriction pathway with *eat-2* mutations[3], to a lesser extent than that of wild type N2 (Supplementary Fig. 2). As the effects of IF and caloric restriction are overlapping, we examined the potential roles of two genes, *skn-1* and *pha-4*, which have been shown to have essential roles in other caloric-restriction regimens, such as dilution of *E. coli* in liquid cultures[4] and *eat-2* mutants[5]. Notably, our IF experiments in a *skn-1*-null mutant, *skn-1(zu135)*, and in *pha-4* RNA interference (RNAi)-subjected animals showed that these genes are dispensable for IF-induced longevity (Supplementary Fig. 2).

In diverse species, TOR activity was downregulated by starvation and inhibition of TOR extended lifespan in a manner similar to

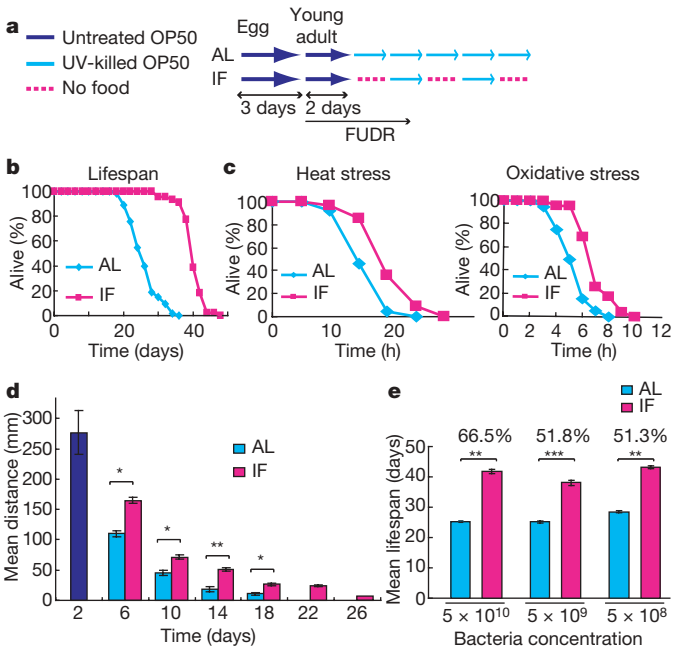


Figure 1 | IF extends *C. elegans* lifespan. **a**, Schematic representation of an IF regimen. AL, *ad libitum*; UV, ultraviolet. **b**, IF increases *C. elegans* lifespan. Survival curves of IF and *ad libitum* worms (wild type N2). **c**, IF increases resistance to heat (left) and oxidative stress (right). Similar results were obtained in two independent experiments. **d**, IF delays the age-dependent decline in locomotion activity. $*P < 0.05$, $**P < 0.01$, *t*-test. **e**, The effects of caloric restriction and IF on lifespan (measured at bacteria concentrations of 5×10^{10}, 5×10^9 and 5×10^8 bacteria ml^{-1}) are additive. $**P < 1.5 \times 10^{-7}$, $***P < 1.0 \times 10^{-11}$, *t*-test. Error bars, s.e.m.

[1]Department of Cell and Developmental Biology, Graduate School of Biostudies, Kyoto University, Sakyo-ku, Kyoto, 606-8502, Japan.

caloric restriction[6–8]. In contrast, RHEB-1, an upstream activator of TOR, was reported to be induced in response to low nutrient levels; the expression of *Drosophila* Rheb was induced by protein starvation and downregulated by subsequent refeeding[9]. These may suggest the complexity of RHEB-1 and TOR signalling responses. We first identified *C. elegans* RHEB-1 as F54C8.5/*rheb-1* (Supplementary Fig. 3a). *rheb-1* RNAi resulted in phenotypes similar to those of TOR-deficient worms (*C. elegans* TOR, B0261.2/*let-363*), such as suppression of endo-reduplication of intestinal nuclei (Supplementary Fig. 3b–e)[10]. These results indicate that F54C8.5 is a *C. elegans* orthologue of Rheb. We then found that from early stages to adulthood, RHEB-1::GFP (a translational fusion of green fluorescent protein to the carboxy terminus of *rheb-1* under the *rheb-1* promoter) is expressed ubiquitously (Supplementary Fig. 3f).

We examined the role of *C. elegans* RHEB-1 and TOR (encoded by the *let-363* gene) in dietary restriction-induced longevity. To downregulate genes of interest, worms were fed on double-stranded RNA (dsRNA)-producing *E. coli* after hatching up until day 2 of adulthood, and adult worms were then subjected to either caloric restriction (Fig. 2a) or IF (Fig. 2b). *rheb-1* RNAi successfully suppressed endogenous RHEB-1 and RHEB-1::GFP expression (Supplementary Fig. 4). Caloric restriction extended the lifespan of control RNAi-treated animals by 18.2% (Fig. 2a, left). Consistent with previous reports, caloric restriction failed to extend the lifespan of TOR (*let-363*) RNAi-treated animals (data not shown). We also found that *rheb-1* RNAi extended lifespan by mimicking the caloric-restriction effects (Fig. 2a, middle). Under the *ad libitum* condition, *rheb-1* RNAi extended lifespan by 19.1%, and the longevity-promoting effect was not seen in the caloric-restriction condition (Fig. 2a, right). Next, we examined the role of the RHEB-1 and TOR pathway in IF (Fig. 2b). IF successfully extended the lifespan of control RNAi-treated animals by 49.3% (Fig. 2b, control RNAi). Surprisingly, we found that inactivation of RHEB-1 and TOR signalling did not mimic the IF effect but suppressed the IF-induced longevity. The mean lifespans of *rheb-1* RNAi-treated *ad libitum* and IF worms were 27.7 and 28.7 days, respectively (Fig. 2b, *rheb-1* RNAi). This clearly indicates the requirement of RHEB-1 for IF-induced longevity. Notably, similarly to *rheb-1* RNAi, TOR (*let-363*) RNAi also suppressed the IF-induced longevity, but the effect of TOR (*let-363*) RNAi was smaller than that of *rheb-1* RNAi; the mean lifespans of TOR (*let-363*) RNAi-treated *ad libitum* and IF worms were 21.6 and 27.5 days, respectively (Fig. 2b, TOR (*let-363*) RNAi). To exclude the possibility that this weak effect of TOR (*let-363*) RNAi is due to insufficient knockdown of TOR, we performed RNAi throughout the whole lifespan using both the feeding and soaking methods (Supplementary Fig. 5). This method of TOR (*let-363*) RNAi also partially suppressed the IF-induced longevity (37.4% lifespan extension by IF), whereas *rheb-1* RNAi almost completely suppressed it (6.0% lifespan extension by IF; Supplementary Fig. 5 and Supplementary Table 1). Therefore, it is unlikely that the partial suppression of IF induced-longevity by TOR (*let-363*) RNAi is due to insufficient knockdown of TOR. To confirm the function of *rheb-1* in adulthood, we performed the soaking RNAi after day 2 of adulthood in IF experiments. Under these conditions, IF increased the lifespan of control RNAi-, *rheb-1* RNAi- and TOR (*let-363*) RNAi-treated worms by 48.4%, 25.3% and 41.3%, respectively, as compared to that under *ad libitum* conditions (Supplementary Fig. 6 and Supplementary Table 1), supporting the notion that RHEB-1 is required for IF-induced longevity. These results indicate that RHEB-1 mediates the IF effects in both TOR-dependent and -independent manners. These results, together with the recent study showing that the absence of food can act as signal independent of calorie intake[11], also suggest that IF is not just an extreme caloric restriction although the caloric restriction and IF effects are overlapping, and that signalling molecules, which mediate caloric restriction and IF stimuli, are different.

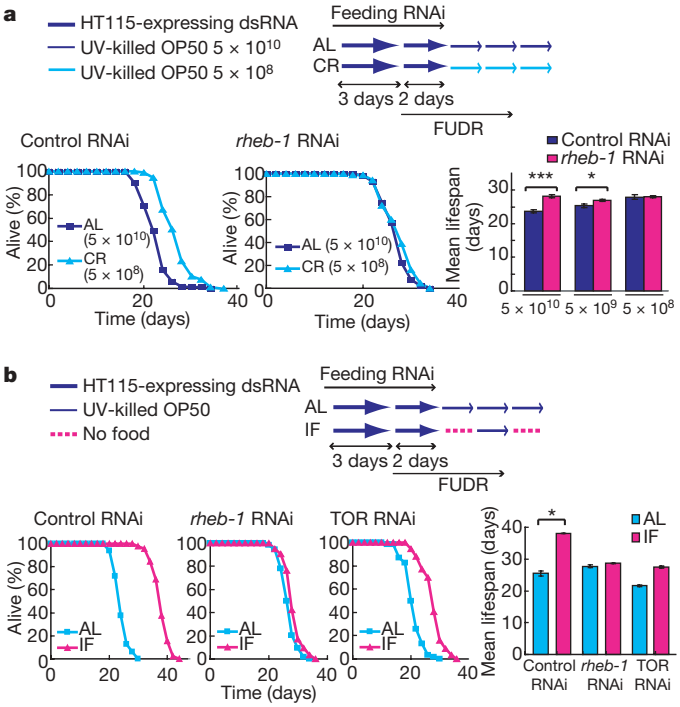


Figure 2 | RHEB-1 has a dual role in dietary restriction. a, *rheb-1* RNAi mimics longevity effects by caloric restriction (CR). Schematic representation of caloric-restriction experiments (top). Survival curves at different concentrations of bacteria (in parentheses) (bottom left and middle). Mean lifespans ± s.e.m. of at least two independent experiments are shown (bottom right, bacterial concentrations shown on the *x*-axis). AL, *ad libitum*; UV, ultraviolet. $*P < 0.05$, $***P < 5 \times 10^{-5}$, *t*-test. **b**, *rheb-1* RNAi and TOR (*let-363*) RNAi suppress longevity effects by IF. Schematic representation of IF experiments (top). Survival curves of control RNAi- (bottom left), *rheb-1* RNAi- (bottom middle) and TOR (*let-363*) RNAi- (bottom right) treated worms in IF. Mean lifespans ± s.e.m. of three independent experiments are shown (bottom far right). $*P < 0.05$, *t*-test.

Because RHEB-1 and TOR signalling functions to promote protein synthesis—the decrease of which increases the lifespan of worms[7,12]—we explored the potential roles of translation-related genes in IF-induced longevity. We performed RNAi for *rps-6* (ribosomal protein subunit 6) in adulthood, and used null mutants for *rsks-1* (p70 S6K) and *ife-2* (translation initiation factor 4E). Both *rsks-1(ok1255)* and *ife-2(ok306)* mutants responded to IF normally (Supplementary Fig. 7). Inactivation of *rps-6* extended lifespan under the *ad libitum* condition, whereas IF further extended the lifespan of *rps-6* RNAi-treated worms to the same extent as that of control RNAi-treated animals by IF (Supplementary Fig. 7 and Supplementary Table 1). These results indicate that decreased translation efficiency cannot account for IF-induced longevity.

We tested several known longevity-regulating and nutrient-sensing genes for their roles in IF. We used their null mutants—*daf-16(mgDf50)*, *daf-16(mu86)*, *aak-2(ok524)*, *sir-2.1(ok434)*, *cep-1(gk138)* and *Y81G3A.3(OK886)* (*C. elegans* GCN2)[13–16]—and found that the IF-induced increase in lifespan was significantly diminished only in two null mutants of *daf-16* ($P = 0.0037$ in *daf-16(mgDf50)*, and $P = 0.0003$ in *daf-16(mu86)*; Fig. 3a and Supplementary Table 1). To examine the possibility that the IF regimen used is not optimized for *daf-16* null mutants, we subjected *daf-16(mu86)* to three kinds of IF regimens: 24 h, 48 h or 72 h of fasting per 4 days. The results showed that all the regimens extended the lifespan of *daf-16(mu86)* to a lesser extent than wild type N2 (Supplementary Fig. 8), showing that *daf-16* partially mediates IF-induced longevity.

DAF-16, the forkhead transcription factor, mediates the effect of the insulin-like signalling pathway on ageing[14]. Environmental stresses, such as starvation, trigger DAF-16 nuclear translocation[17]. We examined whether *rheb-1* RNAi affects fasting-induced nuclear translocation of DAF-16. In control worms, DAF-16::GFP modestly translocated to the nucleus in response to fasting (Fig. 3b). In

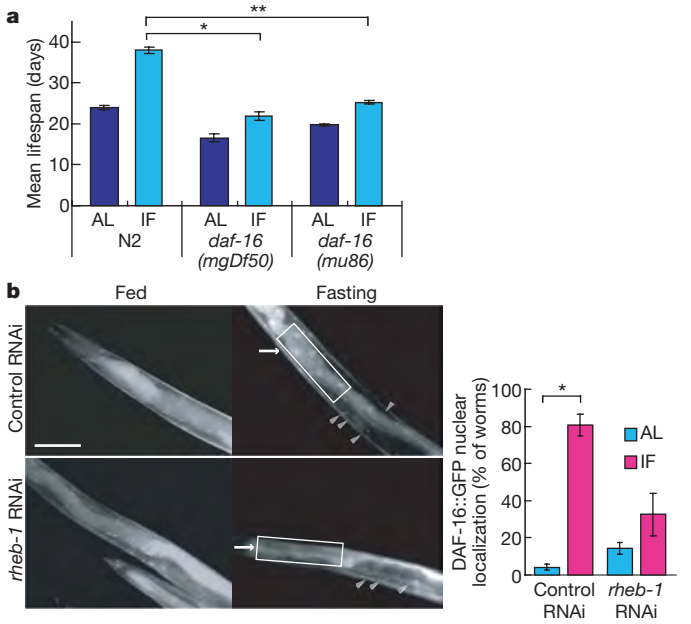


Figure 3 | DAF-16 partially mediates IF-induced longevity. a, Two loss-of-function mutations of *daf-16* suppress IF-induced longevity. Mean lifespans ± s.e.m. of at least three experiments are shown. $*P < 0.001$, $**P < 0.0005$, *t*-test using lifespan extension by IF. AL, *ad libitum*. **b**, Fasting-induced DAF-16::GFP nuclear localization is suppressed by *rheb-1* RNAi in intestine. Representative images are shown (left). Arrows indicate intestine, and white arrowheads indicate other tissues. Scale bar, 100 μm. Percentage of worms with nuclear localization of DAF-16::GFP in intestine after 15 h of fasting were scored in two independent experiments (right). Data are mean ± s.e.m. $*P < 0.05$, *t*-test.

contrast, this fasting-triggered nuclear localization of DAF-16::GFP in the intestine was suppressed by *rheb-1* RNAi (Fig. 3b). Downstream targets of DAF-16—*sod-3*, *mtl-1*, *hil-1* and *dod-6* (ref. 18)—were induced by fasting, and their induction was suppressed by *rheb-1* RNAi and TOR (*let-363*) RNAi (Supplementary Fig. 9, left and middle). Together these findings suggest that DAF-16 mediates, at least in part, the functions of signalling through RHEB-1 in IF-induced longevity. We found that the expression levels of two genes—*rab-10* and *pha-4*, the expression levels of which are reported to be downregulated[19] or upregulated[5], respectively, by caloric restriction—were not affected by fasting (Supplementary Fig. 9, right). Interestingly, the expression level of *pha-4* is upregulated by *rheb-1* RNAi and TOR (*let-363*) RNAi (Supplementary Fig. 9, right), suggesting the possibility that inactivation of RHEB-1 and TOR signalling mimics the caloric-restriction effects through induction of *pha-4* and its downstream targets. It is also reported that the longevity-promoting effect of *pha-4* overexpression is independent of *daf-16*, and *pha-4* is not required for the long lifespan of *daf-2(e1368)*[5]. Therefore, there may be two independent signalling pathways downstream of RHEB-1 and TOR signalling, one of which is the *daf-16* pathway that could mediate fasting-induced longevity.

We next examined the gene expression changes during fasting in *rheb-1* RNAi-treated, TOR (*let-363*) RNAi-treated and control RNAi-treated worms. We performed a genome-wide analysis using Affymetrix GeneChip oligonucleotide microarrays, and compared the gene expression profiles at day 4 of adulthood (see Fig. 4a). We first focused on the effect of fasting, and identified 112 genes which were upregulated by fasting more than threefold with statistical significance in control RNAi-treated worms (see Methods). Surprisingly, the fasting-induced upregulation of most of the 112 genes was dependent on RHEB-1 or TOR (100 genes in RHEB-1 and 94 genes in TOR, Fig. 4b). These 100 RHEB-1-dependent genes can be divided into two classes: genes in which the expression level in *rheb-1* RNAi-fasting is either similar to (59 genes) or less than half (41 genes) that in control-fasting (Fig. 4b). These results show that *rheb-1* RNAi and TOR (*let-363*) RNAi both suppress the induction of these genes by fasting, either by inhibiting fasting-induced upregulation or by inducing upregulation under the fed condition. We then analysed the 112 fasting-induced upregulated genes by scatter plotting (Fig. 4c). The plots of the expression levels in control-fasting versus those in TOR (*let-363*) RNAi-fasting (Fig. 4c, upper left) or those in *rheb-1* RNAi-fasting (Fig. 4c, upper right) show that 24 (yellow) and 42 (red plus yellow) genes are downregulated more than twofold by TOR (*let-363*) RNAi and *rheb-1* RNAi, respectively, and that the 24 genes are included completely in the 42 genes (see also Fig. 4c, lower). This clearly demonstrates that the upregulation of several genes (18 out of 42 genes, 43%) by fasting is dependent on RHEB-1, but not on TOR (Fig. 4c, lower). These results indicate the existence of a TOR-independent pathway, downstream of RHEB-1, which would also mediate IF stimuli to extend lifespan.

We focused on the fasting-induced genes (41 genes, Supplementary Table 2), the induction of which is abolished by *rheb-1* RNAi. We tested several of these for their role in IF-induced longevity by using null mutants (Supplementary Table 1). The loss-of-function mutation of *hsp-12.6*, which encodes a *C. elegans* orthologue of a small heat shock protein αB-crystallin, suppressed the IF-induced increase in lifespan to a similar extent to that in *daf-16(mu86)* (Fig. 4f), suggesting that *hsp-12.6* is one of the downstream targets of DAF-16 in IF-induced longevity. In *daf-16(mu86);hsp-12.6(gk156)*, the extent of IF-induced longevity was similar to that in single mutants *hsp-12.6(gk156)* or *daf-16(mu86)* (Figs 3a, 4f and Supplementary Table 1), confirming that *daf-16* and *hsp-12.6* function in the same signalling pathway. Next, we used long-lived *daf-2* mutants. Low insulin/IGF-like signalling in *daf-2(e1370)* is known to result in constitutive activation of DAF-16 and the higher expression of *hsp-12.6* than in wild type[18,20]. IF did not markedly extend the lifespan of *daf-2(e1370)* mutants (Fig. 4f), suggesting that IF acts by decreasing *daf-2* activity. Furthermore, a reduction-of-function mutation in *hsf-1*, which has been shown to act downstream of *daf-2* to promote longevity by upregulating *hsp-12.6* (ref. 20), also suppressed IF-induced longevity (Supplementary Table 1). Collectively, our data support our idea that reduced *daf-2* signalling, which leads to activation of *daf-16*, *hsf-1* and *hsp-12.6*, mediates the IF effects. Upregulated *hsp-12.6* has been reported to render *daf-2(e1370)* resistant to proteotoxicity, and mutations in αB-crystallin in mammals were reported to cause protein aggregation myopathies[21], suggesting that there is an evolutionarily conserved role for the protein in resistance against proteotoxicity. We found that HSP-12.6::GFP expression is induced by fasting in various tissues including body wall muscles and neuronal systems (Supplementary Fig. 10).

There were 298 genes that were significantly downregulated by fasting more than threefold in control worms. These genes showed decreased expression levels after fasting even in *rheb-1* RNAi- or TOR (*let-363*) RNAi-treated worms (Fig. 4d). Out of the 298 genes, only one, *ins-7*, showed the higher expression level after fasting in *rheb-1* RNAi and TOR (*let-363*) RNAi-treated than in control-fed worms (Fig. 4d, e). It is probable that fasting-induced downregulation of *ins-7*—the finding shown previously[22] and confirmed here (Fig. 4e)—mediates IF-induced longevity, because *ins-7* is shown to negatively regulate longevity by inhibiting *daf-16* activity in a *daf-2*-dependent manner[18]. In addition, in *ZK1251.1e-ins-7(ok1573)*—the strain which carries a null mutation of *ins-7*——the lifespan-extending effect of IF was significantly suppressed (Supplementary Table 1, $P = 0.018$, *t*-test). However, in spite of the marked decrease in the messenger RNA level of *ins-7*, a null deletion of *ins-7* affected IF-induced longevity only modestly. This weak phenotype of *ins-7* deletion may be due to compensation by other insulin-like peptides (*ins* genes). We found that GFP expression under the promoter of one of the *ins* genes, *daf-28*, is markedly suppressed by fasting (Supplementary Fig. 11). As *daf-28* is shown to negatively regulate lifespan[23], these results indicate that only when we could suppress several such

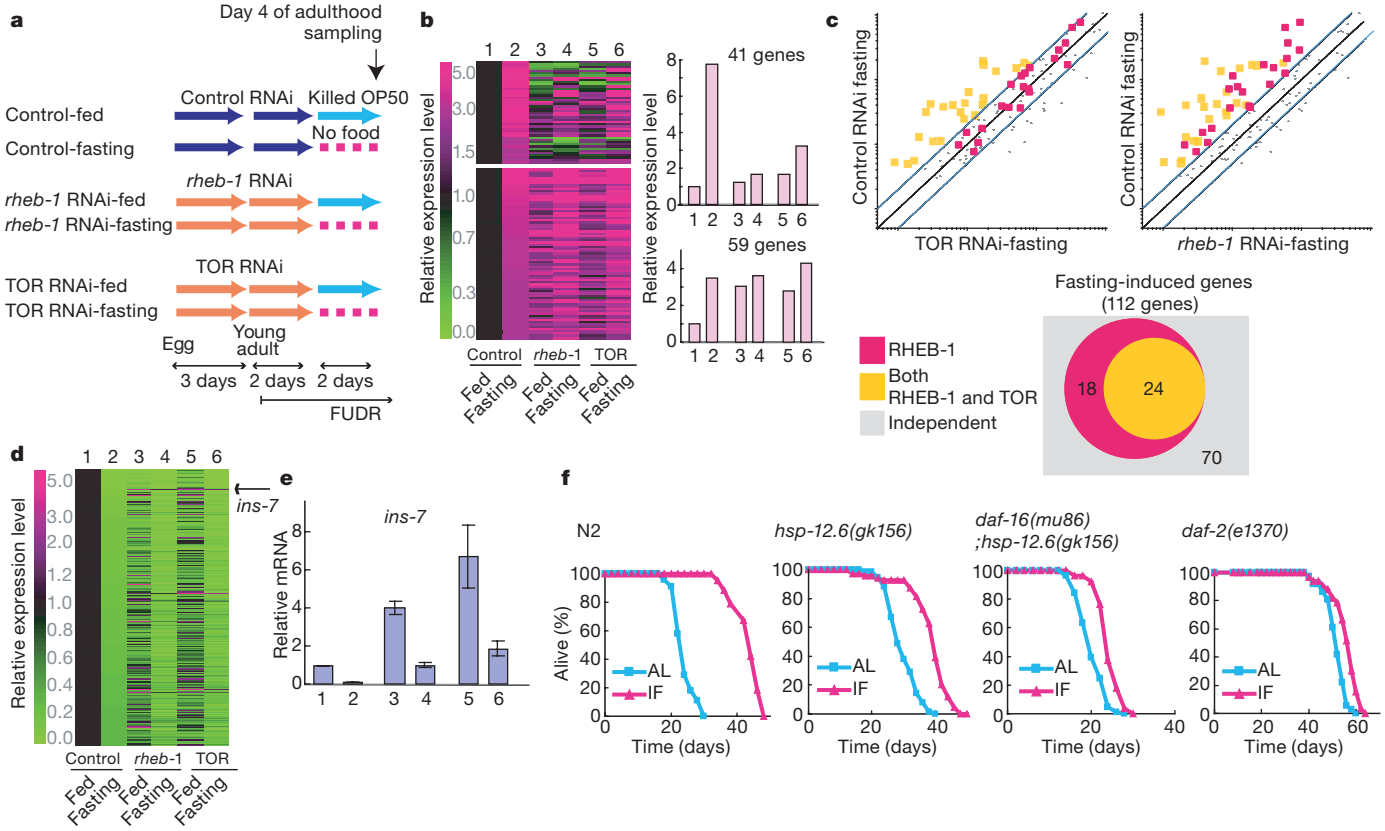


Figure 4 | Microarray analyses identify fasting-regulated genes involved in IF-induced longevity. **a**, The sampling scheme is shown. **b**, **d**, Expression profiles of fasting-induced up-regulated (**b**, left) and down-regulated (**d**) genes. Average expression profiles are shown (**b**, right, see text). **c**, Scatter plots of the expression levels for TOR (*let-363*) RNAi-fasting (top left) or *rheb-1* RNAi-fasting (top right) worms. The black and blue lines indicate the diagonal and twofold changes between two samples. Venn diagram of fasting-induced 112 genes (bottom). **e**, Quantitative PCR with reverse transcription (qRT–PCR) of *ins-7* expression. **f**, HSP-12.6, a downstream target of DAF-16, functions to mediate IF-induced longevity.

agonistic *ins* genes simultaneously, we could appreciably mimic the longevity promoting effect of fasting and markedly affect IF-induced longevity. Interestingly, in *D. melanogaster*, suppressing Tor function increases the level of Dilp2 (also known as Ilp2), an insulin-like peptide[24]. It is thus possible that the regulation of insulin/IGF expression by RHEB-1 and TOR signalling is evolutionarily conserved.

Recent studies have shown that diverse protocols of dietary restriction extend *C. elegans* lifespan through different signalling molecules[4,5]. Dietary food deprivation, in which animals are maintained on plates without food throughout their lives, extends lifespan independently of *daf-2* and *daf-16* signalling[25,26], although this regimen is related to IF. DAF-16 translocates to the nucleus in response to fasting, but relocates to the cytoplasm after prolonged fasting (more than 48 h)[27]. Therefore, worms may have an adaptation system to fasting concerning insulin/IGF-like signalling. Thus, prolonged fasting should be different from IF. A recent study reported that *daf-16* mediates the effect of solid dietary restriction, a kind of caloric restriction[13]. In the study, *aak-2*, a *C. elegans* orthologue of AMPK, is reported to be an upstream regulator of *daf-16*, whereas *aak-2* is dispensable for our IF-induced longevity. Instead, *daf-2* is acting as an upstream regulator of *daf-16* in our study. Thus, signalling pathways mediating IF effects should not be the same as those mediating the other caloric restriction regimen-induced effects. The existence of diverse signalling pathways for so-called 'dietary restriction' may be reasonable, as organisms are exposed to unlimited kinds of stimuli during their lives. Accumulating evidence suggests that these different signalling pathways may converge on the limited number of signalling molecules or physiological reactions. One of these is protein turnover. Autophagy is shown to be essential for longevity in *eat-2* and *daf-2* mutants[28], and reduced translation is shown to extend lifespan by mimicking caloric-restriction effects[7,12]. Bacterial food deprivation as well as lowering insulin/IGF-like signalling confers longevity and resistance to proteotoxicity through *hsf-1* (ref. 29). These findings emphasize the importance of protein quality control. It should be noted that the same pathways or molecules could have different effects on lifespan depending on the ways of dietary restriction or the signalling contexts, as is the case with this study and the previous study in yeast[30]. Further studies are required to clarify the whole picture of signalling networks in dietary restriction. Our results may help the development of dietary restriction mimetics that improve our health without toxic side effects.

METHODS SUMMARY

Lifespan analysis. Approximately 25 young adults per each plate were moved to nematode growth medium (NGM) or RNAi plates containing 200 μg ml^{-1} of 5′fluoro-2′deoxyuridine (FUDR) 3 days after hatching. An adult was scored as dead when it did not respond to a mechanical stimulus. Animals that crawled off the plate, displayed extruded internal organs, or died from internally hatched progeny were censored and excluded from the statistical analysis. All experiments were performed at 20 °C and at least duplicated. *P* values were calculated using a log-rank test or *t*-test with the mean lifespan. See Supplementary Table 1 for mean lifespan, standard error of mean and *P* values of all IF experiments.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 31 July; accepted 23 October 2008.
Published online 14 December 2008.

1. Bordone, L. & Guarente, L. Calorie restriction, SIRT1 and metabolism: understanding longevity. *Nature Rev. Mol. Cell Biol.* **6**, 298–305 (2005).
2. Anson, R. M., Jones, B. & de Cabod, R. The diet restriction paradigm: a brief review of the effects of every-other-day feeding. *Age (Omaha)* **27**, 17–25 (2005).
3. Lakowski, B. & Hekimi, S. The genetics of caloric restriction in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **95**, 13091–13096 (1998).

4. Bishop, N. A. & Guarente, L. Two neurons mediate diet-restriction-induced longevity in *C. elegans*. *Nature* **447,** 545–549 (2007).
5. Panowski, S. H., Wolff, S., Aguilaniu, H., Durieux, J. & Dillin, A. PHA-4/Foxa mediates diet-restriction-induced longevity of *C. elegans*. *Nature* **447,** 550–555 (2007).
6. Kaeberlein, M. *et al.* Regulation of yeast replicative life span by TOR and Sch9 in response to nutrients. *Science* **310,** 1193–1196 (2005).
7. Hansen, M. *et al.* Lifespan extension by conditions that inhibit translation in *Caenorhabditis elegans*. *Aging Cell* **6,** 95–110 (2007).
8. Kapahi, P. *et al.* Regulation of lifespan in *Drosophila* by modulation of genes in the TOR signaling pathway. *Curr. Biol.* **14,** 885–890 (2004).
9. Saucedo, L. J. *et al.* Rheb promotes cell growth as a component of the insulin/TOR signalling network. *Nature Cell Biol.* **5,** 566–571 (2003).
10. Long, X. *et al.* TOR deficiency in *C. elegans* causes developmental arrest and intestinal atrophy by inhibition of mRNA translation. *Curr. Biol.* **12,** 1448–1461 (2002).
11. Smith, E. D. *et al.* Age- and calorie-independent life span extension from dietary restriction by bacterial deprivation in *Caenorhabditis elegans*. *BMC Dev. Biol.* **8,** 49 (2008).
12. Pan, K. Z. *et al.* Inhibition of mRNA translation extends lifespan in *Caenorhabditis elegans*. *Aging Cell* **6,** 111–119 (2007).
13. Greer, E. L. *et al.* An AMPK-FOXO pathway mediates longevity induced by a novel method of dietary restriction in *C. elegans*. *Curr. Biol.* **17,** 1646–1656 (2007).
14. Guarente, L. & Kenyon, C. Genetic pathways that regulate ageing in model organisms. *Nature* **408,** 255–262 (2000).
15. Derry, W. B., Putzke, A. P. & Rothman, J. H. *Caenorhabditis elegans* p53: role in apoptosis, meiosis, and stress resistance. *Science* **294,** 591–595 (2001).
16. Hinnebusch, A. G. Translational regulation of GCN4 and the general amino acid control of yeast. *Annu. Rev. Microbiol.* **59,** 407–450 (2005).
17. Henderson, S. T. & Johnson, T. E. *daf-16* integrates developmental and environmental inputs to mediate aging in the nematode *Caenorhabditis elegans*. *Curr. Biol.* **11,** 1975–1980 (2001).
18. Murphy, C. T. *et al.* Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. *Nature* **424,** 277–283 (2003).
19. Hansen, M., Hsu, A. L., Dillin, A. & Kenyon, C. New genes tied to endocrine, metabolic, and dietary regulation of lifespan from a *Caenorhabditis elegans* genomic RNAi screen. *PLoS Genet.* **1,** e17 (2005).
20. Hsu, A. L., Murphy, C. T. & Kenyon, C. Regulation of aging and age-related disease by DAF-16 and heat-shock factor. *Science* **300,** 1142–1145 (2003).
21. Rajasekaran, N. S. *et al.* Human αB-crystallin mutation causes oxido-reductive stress and protein aggregation cardiomyopathy in mice. *Cell* **130,** 427–439 (2007).
22. Murphy, C. T., Lee, S. J. & Kenyon, C. Tissue entrainment by feedback regulation of insulin gene expression in the endoderm of *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **104,** 19046–19050 (2007).
23. Malone, E. A., Inoue, T. & Thomas, J. H. Genetic analysis of the roles of *daf-28* and *age-1* in regulating *Caenorhabditis elegans* dauer formation. *Genetics* **143,** 1193–1205 (1996).
24. Luong, N. *et al.* Activated FOXO-mediated insulin resistance is blocked by reduction of TOR activity. *Cell Metab.* **4,** 133–142 (2006).
25. Lee, G. D. *et al.* Dietary deprivation extends lifespan in *Caenorhabditis elegans*. *Aging Cell* **5,** 515–524 (2006).
26. Kaeberlein, T. L. *et al.* Lifespan extension in *Caenorhabditis elegans* by complete removal of food. *Aging Cell* **5,** 487–494 (2006).
27. Weinkove, D., Halstead, J. R., Gems, D. & Divecha, N. Long-term starvation and ageing induce AGE-1/PI 3-kinase-dependent translocation of DAF-16/FOXO to the cytoplasm. *BMC Biol.* **4,** 1 (2006).
28. Hansen, M. *et al.* A role for autophagy in the extension of lifespan by dietary restriction in *C. elegans*. *PLoS Genet.* **4,** e24 (2008).
29. Steinkraus, K. A. *et al.* Dietary restriction suppresses proteotoxicity and enhances longevity by an *hsf-1*-dependent mechanism in *Caenorhabditis elegans*. *Aging Cell* **7,** 394–404 (2008).
30. Fabrizio, P. *et al.* Sir2 blocks extreme life-span extension. *Cell* **123,** 655–667 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank members of our laboratory for technical comments and helpful discussion. This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan (to E.N.). Some nematode strains used in this work were provided by the *Caenorhabditis* Genetics Center, which is funded by the NIH National Center for Research Resources (NCRR).

Author Contributions S.H. conceived the study, designed and performed the experiments, and wrote the manuscript with the help of E.N.; S.H. and T.Y. analysed the microarray data; M.U. conducted DAF-16::GFP localization experiments; E.N. supervised the project. All authors discussed the results and commented on the manuscript.

Author Information Microarray data have been deposited with Gene Expression Omnibus at NCBI under the accession number GSE9682. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.N. (L50174@sakura.kudpc.kyoto-u.ac.jp).

METHODS

***C. elegans* strains.** All nematodes were cultured using standard *C. elegans* methods[31]. The strains we analysed were: wild type N2, *sir-2.1(ok434)*(3), *daf-16(mgDf50)*, *daf-16(mu86)*(3), *rsks-1(ok1255)*(2), *rap-2(gk11)*(2), *akk-2(ok524)*(3), *ikb-1(n2027)*, *kin-1(ok338)/mIs13*, *+/szT1*[*lon-2(e678)*] *I*; *kin-2(ok248)/szT1 X*, *ife-2(ok306)*(2), *ZK1251.1&ins-7(ok1573)*(2), *Y81G3A.3(ok886)*(4), *hsp-12.6(gk156)*(4), *cep-1(gk138)*(2), *clk-1(e2519)*(2), *skn-1(zu135)*(3), *hsf-1(sy441)*(2), *ccls4251*(1), *hil-1(gk229)*(2), GR1455 *mgls40*[*Pdaf-28*::gfp] and TJ356 *zls356*[*daf-16*::gfp; *rol-6*](2). The numbers of outcrossing are shown in parentheses.

Intermittent fasting. Approximately 100 synchronized young adult worms raised on NGM plates with live OP50 were picked to FUDR-containing plates with live OP50. At day 2 of adulthood, worms were divided into *ad libitum* and IF. Worms in *ad libitum* were fed ultraviolet-killed OP50 *ad libitum* throughout their lifespan. Worms in IF were on plates with ultraviolet-killed OP50 or without food alternatively every other day. All worms were transferred to new plates every other day.

Solid dietary restriction. Solid dietary restriction was performed as described[13] with modifications. The OP50 concentration was calculated by counting colony performing units. OP50 was resuspended and diluted in water. One-hundred-and-fifty microlitres of these solutions were plated and bacteria were immediately killed by ultraviolet irradiation.

RNA interference. RNAi was performed by the feeding method[32] and/or the soaking method[33] as described. The first 500 nucleotides of the coding region of F54C8.5/*rheb-1*, TOR/*let-363a*, *rps-6* and *pha-4*c complementary DNA were used for RNAi. The primers used were: *rheb-1* forward, 5′-ATGAGCAGTTCGCTGCAA-3′; *rheb-1* reverse, 5′-AACACCTCATGCACTCGA-3′; TOR (*let-363*) forward, 5′-ATGCTCCAACAACACGGA-3′; TOR (*let-363*) reverse, 5′-GCTTTTGAAGCCATCTTG-3′; *rps-6* forward, 5′-ATGAGACTTAACTTCGCC-3′; *rps-6* reverse, 5′-CCATCTGGGAAGGTCTTG-3′; *pha-4* forward, 5′-ATGAACGCTCAGGACTATCT-3′; *pha-4* reverse, 5′-CAATTGACATTGTCTGAAAT-3′. Primers were fused with restriction enzyme sites (for feeding RNAi) or T3 and T7 promoter sequences (for soaking RNAi). In brief, in feeding RNAi, each cDNA segment was cloned into the feeding vector pPD129.36 with a Sac2 and a Kpn1 site and transformed to HT115 bacterial cells. Control animals were fed bacteria carrying an empty pPD129.36 vector. In soaking RNAi, cDNA segments were fused with T3 and T7 promoter sequences and RNA was synthesized by T3 and T7 RNA polymerase. Worms were soaked in soaking buffer containing 1–1.5 mg ml^{-1} dsRNA. Control animals were soaked with soaking media without dsRNA.

Generation of transgenic animals. To generate GFP reporter constructs, fragments containing the 5′ upstream sequence and coding region were amplified using PCR. The sizes of fragments were: F54C8.5 2.2 kilobase (kb), *hsp-12.6-1kbp* 1.6 kb and *hsp-12.6-5kbp* 5.5 kb. These fragments were cloned into the GFP vector pPD95.75. Transgenic worms were generated by microinjecting these plasmids into wild type N2 worms at 20 ng μl^{-1}, 20 ng μl^{-1} and 50 ng μl^{-1}, with 50 ng μl^{-1} pRF4 *rol-6* transformation marker.

qRT–PCR. Total RNA was extracted with Sepasol(R)-RNA ISuper (Nacalai tesque), purified with RNeasy Mini Kit (Qiagen), and reverse transcribed into cDNA using M-MLV reverse transcriptase (Invitrogen) with dT primer, according to manufacturers' instructions. cDNA was subjected to qPCR analysis using the ABI 7300Real Time PCR System (Applied Biosystems) with SYBR GreenPCR Kit (Roche). Each value was normalized to *act-1*, and the value in control RNAi-fed was set to 1. Primer sequences are available on request.

Heat and oxidative stress resistance. Synchronous worms at 10 day of adulthood were washed with M9 buffer and resuspended in M9 buffer with 20 animals per well. At time 0, hydrogen peroxide was added to the well in a final concentration 2.5 mM. The surviving worms were scored every hour. Heat stress was given by incubating worms on NGM plates at 35 °C. The surviving worms were scored every 5 h.

Locomotion activity. The length of worm tracks in 1 h was photographed and measured using the software application Axioplan2.

Fluorescent microscopy. For intestinal lysosomes, animals were anaesthetized with 10 mM sodium azide in M9 buffer at 120 h after synchronization and observed with Axioplan2. For intestinal nuclei, 4-day-old adult worms were fixed with 75% methanol in PBS for 1 h, washed twice with PBS and incubated with 0.1% Hoechst. For GFP::NLS in body wall muscles, the worms carrying the transgene *ccls4251* were fixed with 3% formaldehyde in PBS for 5 min at room temperature. For HSP-12.6::GFP expressing strains, worms were anaesthetized with 10 mM sodium azide in M9 buffer at 4 day of adulthood. In DAF-16 localization assay, worms expressing DAF-16::GFP were synchronized and grown in the following conditions: control fed, control fasting, *rheb-1* RNAi-fed and *rheb-1* RNAi-fasting. After 15 h fasting, worms were fixed with 3% formaldehyde in PBS for 5 min at room temperature.

Body size. Animals were synchronized by breaching gravid hermaphrodites. One-hundred-and-twenty hours after synchronization, worms were anaesthetized with 10 mM sodium azide in M9 buffer and photographed using Axioplan2. To measure body length and width, segmented lines were drawn and the length of lines was calculated using the software application Axioplan2.

Microarray experiments. Five, four and three independent experiments were performed in control RNAi, *rheb-1* RNAi and TOR (*let-363*) RNAi, respectively (see Fig. 4a). Total RNA was extracted as described above. Other procedures were performed according to Affymetrix protocols. Hybridized arrays were scanned using an Affymetrix GeneChip Scanner. Scanned chip images were analysed with GeneChip Operating Software v.1.4 (GCOS), and processed using default settings. The Affymetrix output (CEL files) was imported into GeneSpring 7.3 (Agilent Technologies) microarray analysis software for both statistical analysis and presentation of expression profiles (average expression profiles and scatter plots). Expression signals of probe sets were calculated using GCRMA (GC robust multi-array analysis, as implemented in GeneSpring). The log of ratio mode was used for all analyses (GeneSpring). The data have been submitted to the GEO at NCBI under the accession number GSE9682. Statistical analysis was performed by one-way analysis of variance (ANOVA) with a Benjamini and Hochberg false discovery rate (BH-FDR = 0.1) multiple testing correction followed by Tukey post-hoc tests using log-transformed data (GeneSpring). Three-thousand-and-thirty-one probe sets were identified to be regulated by fasting in control RNAi-treated worms. We defined RHEB-1-dependent or TOR-dependent genes as those in which the expression level induced by fasting was reduced to less than half under *rheb-1* RNAi or TOR (*let-363*) RNAi conditions.

31. Brenner, S. The genetics of *Caenorhabditis elegans. Genetics* **77**, 71–94 (1974).
32. Kamath, R. S., Martinez-Campos, M., Zipperlen, P., Fraser, A. G. & Ahringer, J. Effectiveness of specific RNA-mediated interference through ingested double-stranded RNA in *Caenorhabditis elegans*. *Genome Biol.* **2**, research0002.1–research0002.10 (2001).
33. Maeda, I., Kohara, Y., Yamamoto, M. & Sugimoto, A. Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Curr. Biol.* **11**, 171–176 (2001).

LETTERS

Chlamydia causes fragmentation of the Golgi compartment to ensure reproduction

Dagmar Heuer[1], Anette Rejman Lipinski[1], Nikolaus Machuy[1], Alexander Karlas[1], Andrea Wehrens[1], Frank Siedler[3], Volker Brinkmann[2] & Thomas F. Meyer[1]

The obligate intracellular bacterium *Chlamydia trachomatis* survives and replicates within a membrane-bound vacuole, termed the inclusion, which intercepts host exocytic pathways to obtain nutrients[1–3]. Like many other intracellular pathogens, *C. trachomatis* has a marked requirement for host cell lipids, such as sphingolipids and cholesterol, produced in the endoplasmic reticulum and the Golgi apparatus[4–6]. However, the mechanisms by which intracellular pathogens acquire host cell lipids are not well understood[1–3]. In particular, no host cell protein responsible for transporting Golgi-derived lipids to the chlamydial inclusions has yet been identified. Here we show that *Chlamydia* infection in human epithelial cells induces Golgi fragmentation to generate Golgi ministacks surrounding the bacterial inclusion. Ministack formation is triggered by the proteolytic cleavage of the Golgi matrix protein golgin-84. Inhibition of golgin-84 truncation prevents Golgi fragmentation, causing a block in lipid acquisition and maturation of *C. trachomatis*. Golgi fragmentation by means of RNA-interference-mediated knockdown of distinct Golgi matrix proteins before infection enhances bacterial maturation. Our data functionally connect bacteria-induced golgin-84 cleavage, Golgi ministack formation, lipid acquisition and intracellular pathogen growth. We show that *C. trachomatis* subverts the structure and function of an entire host cell organelle for its own advantage.

In the Golgi apparatus newly synthesized proteins and lipids, including sphingolipids, are modified in a stepwise process and then sorted to various locations inside, or are exported out of the cell[7]. The organization of the Golgi apparatus is thought to depend on cytoplasmic structural proteins, including golgins, which form the Golgi matrix[8]. Golgins belong to a large family of proteins with coiled-coil domains that localize to the Golgi apparatus[9]. Fragmentation of the Golgi apparatus is a common feature of a number of physiological processes such as mitosis and apoptosis. During mitosis, fragmentation is thought to be triggered by the phosphorylation of several proteins, including golgin-84, GM130 and GRASP-65 as well as the inactivation of small GTPases[10–13]. During apoptosis, fragmentation is induced by caspase-dependent cleavage of golgins, such as giantin, golgin-160 and p115 (refs 14–16).

To investigate the mechanisms underlying sphingolipid trafficking to the chlamydial inclusion, we began by observing Golgi apparatus structure in *C. trachomatis*-infected epithelial cells. Live-cell microscopy revealed a focused GFP–GM130 signal, typical of the normal ribbon-like structure of the Golgi apparatus, in close association with the bacterial inclusion at an early stage of infection (Fig. 1a and Supplementary Movie 1). As infection proceeded the signal expanded, indicating a progressive dispersion of GFP–GM130-positive structures, resulting in the disassembly of the Golgi ribbon structure. The remaining smaller Golgi elements, albeit increased in number (Fig. 1b and Supplementary Fig. 1), were aligned along the inclusion membrane (Fig. 1a, c). Transmission electron microscopy (TEM) revealed a typical Golgi structure consisting of laterally linked stacked cisternae extending over several micrometres in non-infected cells (Fig. 1d), whereas in infected cells a fragmented Golgi apparatus composed of relatively short Golgi stacks that were neither laterally linked nor aligned to each other was observed (Fig. 1e). Our results indicate that *C. trachomatis* infection induces fragmentation of the Golgi apparatus into small, albeit intact, Golgi ministacks.

We then investigated the status of various golgins in infected cells. Immunoblotting of lysates revealed that golgin-84 was sequentially processed during infection, yielding two distinct fragments of ~78 kDa and ~65 kDa; the smaller fragment accumulated at later time points after infection (Fig. 2a). Cleavage was dependent on time and multiplicity of infection (MOI; Fig. 2a and Supplementary Fig. 2). Notably, golgin-84 cleavage also occurred in a range of epithelial cell lines infected with a variety of chlamydial strains (Supplementary Figs 3 and 4). Thus, infection with *Chlamydia* species leads to stepwise processing of golgin-84, accompanied by Golgi fragmentation.

Cleavage of golgins has previously been associated with Golgi fragmentation during apoptosis[14–16]. Although *Chlamydia*-infected cells are largely protected from apoptosis[17,18], we investigated the possible involvement of caspases in golgin-84 cleavage. Treatment of infected cells with pan-caspase inhibitor IV and Z-WEHD-FMK, an inhibitor of inflammatory caspases, elicited a near-complete blockage of *C. trachomatis*-induced cleavage of golgin-84. In contrast, cleavage was unaffected by the addition of the apoptotic caspase inhibitor Z-DEVD-FMK (Fig. 2b). Furthermore, stable expression of various amino-terminally truncated versions of golgin-84 in golgin-84 knockdown cells indicated that the golgin-84 cleavage site leading to the 65-kDa fragment is most likely contained within amino acids 148–158 (Supplementary Fig. 5). To determine the precise cleavage site generating the 65-kDa fragment, golgin-84 fused to Myc tag was transiently expressed in *Chlamydia*-infected cells. Tagged golgin-84 was precipitated from cells late in the infection cycle to assure complete cleavage of golgin-84 into the predominant 65-kDa fragment. Mass spectroscopy analysis of isolated and trypsin-digested golgin-84 demonstrated that a major proportion of golgin-84 was cleaved at amino acid S157 (Supplementary Fig. 6). *In silico* analysis of amino acids 155–157 (http://merops.sanger.ac.uk/) revealed calpain 2 as a candidate protease for golgin-84 cleavage. Calpains represent a group of intracellular cysteine proteases that locate to the cytosol, endoplasmic reticulum and Golgi apparatus[19,20] and are considered as biomodulators[21]. To test their presumptive role, infected cells were treated with several specific membrane-permeable inhibitors. Both calpain inhibitors, calpeptin and calpain inhibitor III (CalpIII), nearly completely blocked the generation of the 65-kDa golgin-84 fragment in

[1]Department of Molecular Biology, [2]Microscopy Core Facility, Max Planck Institute for Infection Biology, Charitéplatz 1, 10117 Berlin, Germany. [3]Department of Membrane Biochemistry, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany.

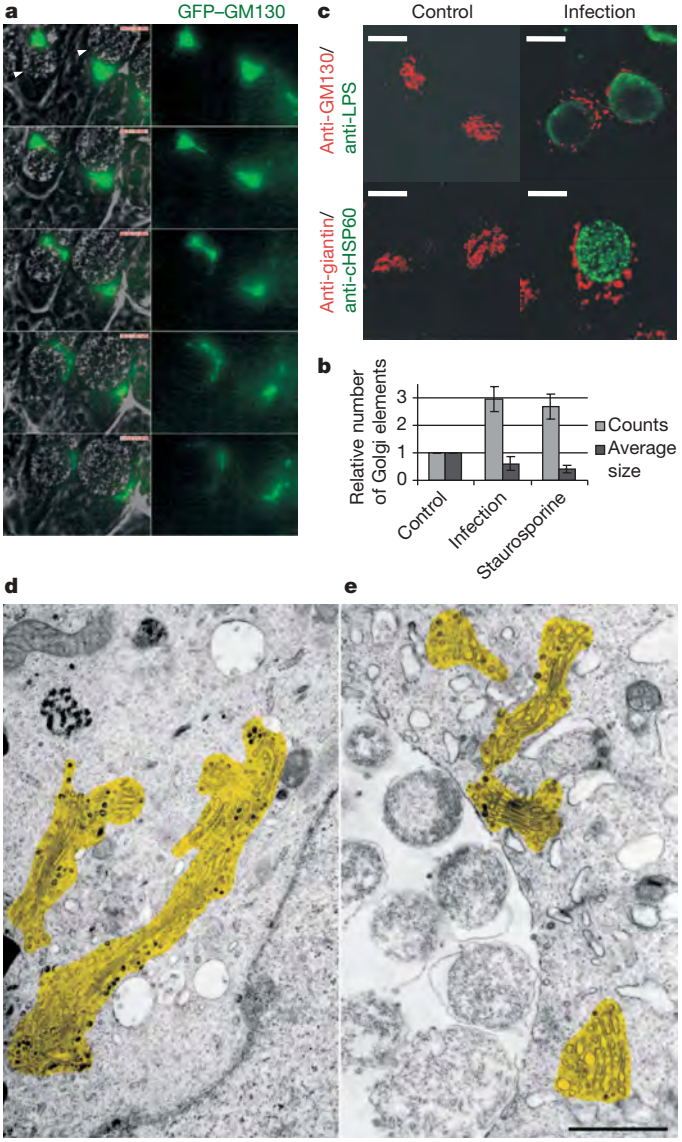


Figure 1 | *Chlamydia trachomatis* infection triggers breakdown of Golgi apparatus into ministacks. **a**, Time-lapse microscopy of GFP–GM130-expressing HeLa cells infected with *C. trachomatis* (MOI = 2). Arrowheads indicate inclusions. Original magnification, ×283. **b**, Comparison of numbers and average size of GM130 signal in control, infected and staurosporine-treated cells. Data show mean ± s.d. of three independent experiments. **c**, Control and infected HeLa cells were stained with antibodies against GM130 or giantin (red channel) 28 h after infection. *Chlamydia* was detected using antibodies to lipopolysaccharide (LPS) or HSP60 (green channel). cHSP60, chlamydial HSP60. Scale bar, 10 μm. **d, e**, Ultrastructural analysis of control (**d**) and cells infected with *C. trachomatis* for 24 h (**e**). Yellow area, Golgi apparatus. Scale bar, 1 μm.

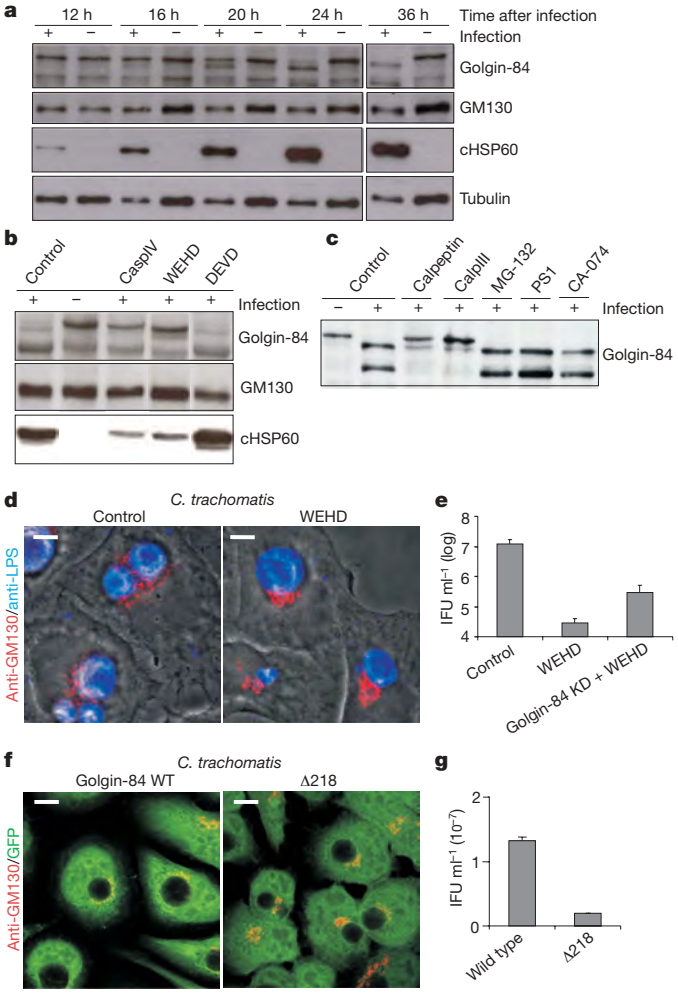


Figure 2 | Cleavage of golgin-84 in infected cells is associated with Golgi fragmentation. **a**, Immunoblots of lysates (hours after infection indicated) showing golgin-84, GM130, giantin, cHSP60 and tubulin expression in infected (+) or mock-infected HeLa cells (−). **b**, Immunoblots of lysates from *C. trachomatis*- (+) or mock-infected (−) HeLa cells (24 h after infection), after addition of caspase inhibitors Z-WEHD-FMK (caspase-1/5 inhibitor), Z-DEVD-FMK (caspase-3/7 inhibitor) or caspase inhibitor IV (pan-caspase inhibitor) at 9 h after infection. **c**, Immunoblots of lysates from *C. trachomatis*- (+) or mock-infected (−) HeLa cells (26 h after infection) treated with specific calpain (calpeptin, CalpIII), proteasome (MG-132, PS1) and cathepsin (CA-074) inhibitors at 8 h after infection. **d**, Merge of GM130 (red channel) and LPS (blue channel) in untreated (control) or Z-WEHD-FMK-treated cells after infection. Scale bar, 10 μm. **e**, *C. trachomatis* maturation in untreated control, WEHD-FMK-treated control and golgin-84 knockdown (KD) cells 48 h after infection. Numbers of infectious bacteria measured as IFU ml^{-1} are depicted in log scale. Data show mean + s.d. of duplicates. **f**, GM130 immunostaining of infected golgin-84 knockdown cells stably expressing wild-type (WT) golgin-84 or the mutant (Δ218). Inclusions are seen as black holes in the GFP channel, as GFP does not cross the inclusion membrane. Scale bar, 10 μm. **g**, Replication of *C. trachomatis* in mutant (Δ218) cell lines at 48 h after infection. Numbers of infectious bacteria are measured as total IFU ml^{-1}. Data show mean + s.d. of duplicates.

infected cells, whereas formation of the 78-kDa fragment was less affected (Fig. 2c). Taken together, our data reveal that both inflammatory caspases and calpains are involved in the successive cleavage of golgin-84 and that the 65-kDa fragment of golgin-84 is probably generated by calpain cleavage at position S157.

We next assessed the interdependence between golgin-84 cleavage, Golgi fragmentation and chlamydial propagation. Golgin-84 cleavage was blocked via treatment of infected cells with Z-WEHD-FMK, resulting in a lack of Golgi fragmentation (Fig. 2d and Supplementary Fig. 7) and a 2-log reduction in numbers of infectious bacteria (Fig. 2e). Bacterial numbers decreased until six days after infection, indicating a profound block in bacterial maturation (Supplementary Fig. 8). RNA interference (RNAi) was then used to test the inhibitory effect of Z-WEHD-FMK on chlamydial replication in the absence of golgin-84. Notably, short interfering RNA (siRNA) knockdown of golgin-84 restored chlamydial growth in host cells treated with Z-WEHD-FMK (Fig. 2e). To further support our hypothesis that golgin-84 can directly influence Golgi structure and *Chlamydia* replication, we generated an N-terminal deletion mutant (Δ218) of golgin-84 lacking potential protein interaction sites and inhibiting *Chlamydia*-induced fragmentation of the Golgi apparatus (Fig. 2f). Stable cell lines expressing either wild-type or Δ218 golgin-84 were generated by lentiviral transduction of golgin-84 knockdown

cells. Numbers of infectious bacteria were reduced by ~5 times in cells expressing Δ218 golgin-84 (Fig. 2g). Taken together, these data reveal that Golgi fragmentation in infected cells is a downstream event of golgin-84 cleavage and that Golgi fragmentation is a critical factor for efficient chlamydial growth.

We hypothesized that Golgi fragmentation before infection could boost bacterial replication. Therefore, Golgi fragmentation was induced by knockdown of giantin, GPP130 (albeit less prominently) and golgin-84 (ref. 11; see Fig. 3a and Supplementary Figs 9–13). These siRNA and short hairpin RNA (shRNA) golgin-84 knockdown cells, plus luciferase knockdown control cells, were then infected with *C. trachomatis* and the infectious progeny quantified at various times after infection. By 48 h after infection up to six times more infectious bacteria were found in siRNA golgin-84 and giantin knockdown cells and up to three times more in GPP130 knockdown cells (Fig. 3b), compared with control cells. In shRNA golgin-84 knockdown cells about three times more infectious bacteria could be recovered as early as 24 h after infection compared with control cells, increasing to nearly ten times more infectious bacteria at 40 h after infection (Fig. 3c). Furthermore, treatment of infected cells at 8 h after infection with a very low dose of nocodazole, still sufficient to induce Golgi fragmentation, enhanced bacterial propagation (Supplementary Fig. 14). Thus, Golgi fragmentation by various means boosts *C. trachomatis* reproduction. Interestingly, *Salmonella enterica* serovar Typhimurium, a facultative intracellular bacterium, did not show enhanced replication in golgin-84 knockdown cells (Supplementary Fig. 15), indicating a neutral function for golgin-84 in *Salmonella* infections. On the basis of the fact that no difference in numbers of inclusions in infected golgin-84 knockdown cells compared with control cells at low MOI (Supplementary Fig. 16) was observed, we excluded the possibility that the growth stimulation elicited in golgin-84 shRNA cells was due to an increase in the primary infection. Consistent with the hypothesis of accelerated replication in golgin-84 knockdown cells, electron microscopy revealed the development of small, electron-dense particles, indicative of infectious bacteria (elementary bodies; see Methods), as early as 24 h after infection (Fig. 3d), but not in infected control cells (Fig. 3e). Taken together, these results support the notion that golgin-84 inactivation has a decisive role in growth regulation of *C. trachomatis*, in that chlamydial maturation is dependent on, and can be enhanced by, depletion of golgin-84 throughout the replication cycle.

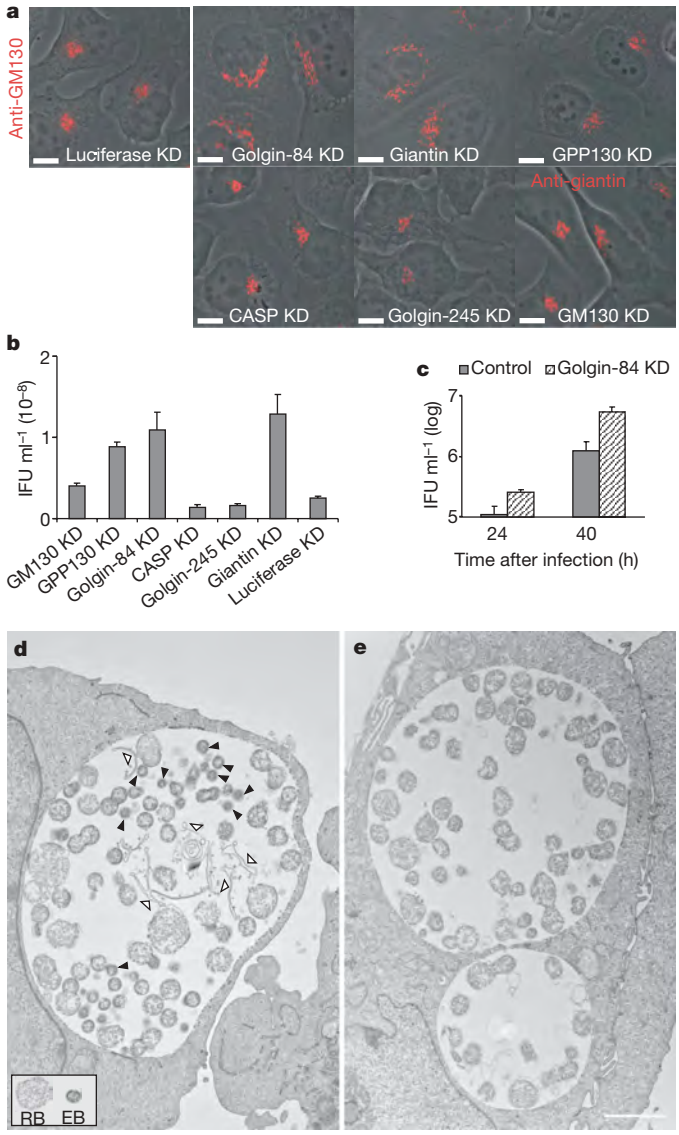


Figure 3 | RNAi-mediated fragmentation of the Golgi apparatus enhances chlamydial propagation. **a**, GM130 immunostaining of HeLa cells transfected with siRNAs directed against the indicated golgins four days after transfection. GM130 knockdown (KD) cells were immunostained for giantin to visualize the Golgi apparatus. Scale bar, 10 μm. **b**, **c**, Numbers of infectious bacteria measured as IFU ml^{-1} at 48 h after infection in various transient golgin knockdown cells (**b**) and in stable golgin-84 knockdown and control cell lines at 24 h and 40 h after infection (**c**; in log scale). Representative experiments performed in duplicates are shown. Data show mean + s.d. of duplicates. **d**, **e**, TEM of *C. trachomatis*-infected human epithelial cells (24 h after infection). **d**, **e**, Stable golgin-84 knockdown (**d**) and control cells (**e**). Black arrowheads, infectious bacteria (elementary bodies, EB); white arrowheads, membranous structures. A representative picture of a non-infectious reticular body (RB) and an infectious elementary body (EB) is depicted in the inlay. Scale bar, 2 μm.

Finally, we reasoned that a breakdown of the Golgi structure into more, albeit smaller, Golgi elements aligned around the inclusion may enhance lipid transport. Therefore, we treated infected cells with Z-WEHD-FMK, effectively preventing Golgi fragmentation, or with DMSO as a control and then labelled cells with fluorescent ceramide. Confocal images revealed that ceramide was rapidly incorporated into the inclusion membrane within DMSO-treated cells and accumulated inside the inclusion in bacterial membranes (Fig. 4b and Supplementary Movie 2). In contrast, Z-WEHD-FMK-treated cells were only slightly fluorescent as the majority of lipid accumulated in a Golgi-like structure outside the inclusion (Fig. 4b and Supplementary Movie 3). These observations clearly show that Golgi fragmentation enhances transport of sphingolipids to the bacterial inclusion at later time points during the infection.

Here we show that infection of human epithelial cells with *C. trachomatis* induces a fragmentation of the Golgi apparatus triggered by the successive cleavage of golgin-84, which is affected by inhibitors of both inflammatory caspases and calpains. Yet, this does not exclude a role for bacterial proteases in golgin-84 processing. Truncation of golgin-84 resulted in the formation of Golgi ministacks around the bacterial inclusion. Inhibition of golgin-84 cleavage or expression of an inhibitory golgin-84 mutant (Δ218) prevented Golgi breakdown. Inhibition of Golgi fragmentation substantially

As Golgi apparatus structural alterations resulting from GM130 depletion are reported to cause an accumulation of improperly glycosylated proteins in the plasma membrane[22], we reasoned that *Chlamydia*-induced Golgi fragmentation could also interfere with processing of glycoproteins. As a marker of premature glycoproteins, proteins with terminal *N*-acetyl-D-glucosamine on the plasma membrane were detected using the GS-II lectin from *Griffonia simplicifolia*[22]. High levels of lectin binding were indicated by a bright staining of the plasma membrane in infected cells, whereas in uninfected cells no intense staining was observed (Fig. 4a). Although the Golgi apparatus still delivered glycoproteins to the cell surface, the normal processing of glycoproteins was altered in infected cells. Thus, *C. trachomatis*-induced Golgi fragmentation affects the processing of glycoproteins in the Golgi apparatus.

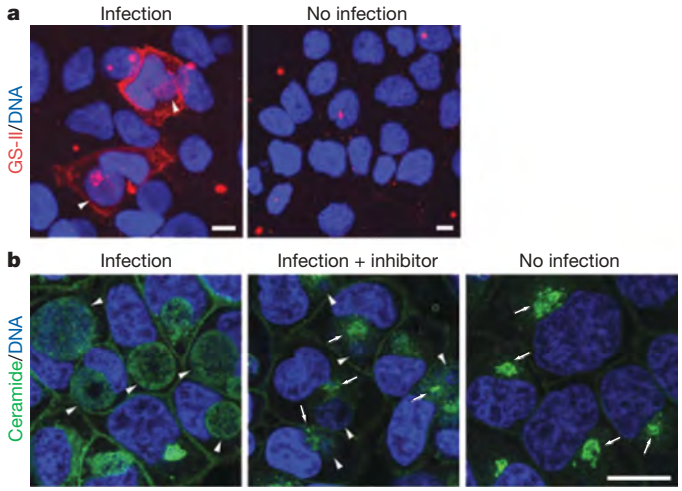


Figure 4 | Functions of a fragmented Golgi apparatus. a, Infected or uninfected cells were stained with GS-II Alexa Fluor 594 28 h after infection. Nuclei and bacteria were counterstained with Hoechst. Scale bar, 10 μm. **b**, HeLa cells were either left uninfected (no infection) or infected with *C. trachomatis* (MOI = 2). One infected sample was treated with 80 μM Z-WEHD-FMK 9 h after infection (infection + inhibitor). BODIPY FL C5-ceramide transport in infected cells with or without inhibitor and in control cells was analysed by time-lapse microscopy 30 min after addition of labelled ceramide. Nuclei and bacteria were counterstained with Hoechst. Scale bar, 10 μm; arrowheads, inclusions; arrows, Golgi apparatus.

reduced transport of lipids to the *C. trachomatis* inclusion and severely blocked bacterial propagation. Thus, golgin-84 turns out to be a crucial modulator of both structure and function of the Golgi apparatus during *Chlamydia* infection. Stable knockdown of golgin-84 (and also giantin) resulted in fully viable cells showing a fragmented Golgi apparatus, leading to a substantial enhancement of chlamydial development.

Earlier work has indicated that sphingolipid acquisition begins as early as 2 h after infection, when internalized elementary bodies have been converted into metabolically active reticular bodies (see Methods)[2,3]. However, the requirement for sphingolipids is initially low and thought to increase markedly with increased bacterial replication and expansion of the inclusion. This advanced stage of chlamydial expansion, starting at ~20 h after infection, coincides with Golgi fragmentation (Supplementary Fig. 17). Notably, different molecular mechanisms of lipid transfer could account for the early and advanced lipid acquisition. For instance, whole Golgi-derived vesicles might fuse with the inclusion membrane[3]. Alternatively, lipids might be transported individually to the inclusion membrane via specific lipid transporters or by pathways that bypass the Golgi apparatus[23,24]. It will be of interest to determine which of these transport mechanisms is used in the early and/or advanced stages of chlamydial development.

Our discovery reveals a novel infectious mechanism by which an intracellular pathogen morphologically and functionally manipulates a host organelle to enhance lipid acquisition and secure its replication and development. Moreover, this work has identified novel molecular targets, including inflammatory caspases and calpains, which may potentially prove useful in the treatment of chlamydial infections.

METHODS SUMMARY

Infection and quantification of *Chlamydia* progeny. *Chlamydia trachomatis* serovar LGV L2, A and K and *Chlamydia muridarum* were propagated in HeLa cells in medium with 5% FCS at 35 °C in 5% CO_2. For infection, elementary bodies were adsorbed to HeLa cells with a multiplicity from 0.5 to 5 and incubated for varying times, depending on assay. Numbers of chlamydial progeny were measured as the number of inclusion-forming units (IFU) by counting inclusions on a fluorescence microscope using a ×40 objective. IFUs were expressed as IFU ml^{-1}.

Immunofluorescence and microscopy. Cells were seeded onto coverslips and treated as indicated. Cells were then fixed with 2% PFA for 30 min at room temperature. Different Golgi markers were detected using specific antibodies and bacteria were detected in infected cells using either rabbit anti-lipopolysaccharide or mouse anti-HSP60 antibodies, followed by specific secondary fluorescently labelled antibodies and mounted in MOWIOL. Images were taken with a Leica TCS-SP confocal microscope and processed using Adobe Photoshop 6.0.

siRNA. All siRNAs were designed and purchased from Qiagen. siRNA validation was performed according to ref. 25. HeLa cells were transfected using RNAiFect (Qiagen) according to the manufacturer's guidelines.

Live-cell imaging of ceramide transport. Cells were seeded onto glass-bottom dishes (MatTek Corporation) and infected with *C. trachomatis* (MOI = 2) or left uninfected for control. At indicated time points after infection, the dishes were transferred to an inverted confocal microscope (SP5, Leica) equipped with an incubator heated to 37 °C. Every 30 s, a *Z*-stack of 20 frames covering a depth of 10 μm was recorded for fluorescence and DIC using a resonant scanner. BODIPY FL C5-ceramide was added directly to the cells and bacterial nuclei were stained using Hoechst.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 22 July; accepted 22 October 2008.
Published online 7 December 2008.

1. Carabeo, R. A., Mead, D. J. & Hackstadt, T. Golgi-dependent transport of cholesterol to the *Chlamydia trachomatis* inclusion. *Proc. Natl Acad. Sci. USA* **100,** 6771–6776 (2003).
2. Hackstadt, T., Rockey, D. D., Heinzen, R. A. & Scidmore, M. A. *Chlamydia trachomatis* interrupts an exocytic pathway to acquire endogenously synthesized sphingomyelin in transit from the Golgi apparatus to the plasma membrane. *EMBO J.* **15,** 964–977 (1996).
3. Scidmore, M. A., Fischer, E. R. & Hackstadt, T. Sphingolipids and glycoproteins are differentially trafficked to the *Chlamydia trachomatis* inclusion. *J. Cell Biol.* **134,** 363–374 (1996).
4. Hatch, G. M. & McClarty, G. Phospholipid composition of purified *Chlamydia trachomatis* mimics that of the eucaryotic host cell. *Infect. Immun.* **66,** 3727–3735 (1998).
5. van Ooij, C. *et al.* Host cell-derived sphingolipids are required for the intracellular growth of *Chlamydia trachomatis*. *Cell. Microbiol.* **2,** 627–637 (2000).
6. Wylie, J. L., Hatch, G. M. & McClarty, G. Host cell phospholipids are trafficked to and then modified by *Chlamydia trachomatis*. *J. Bacteriol.* **179,** 7233–7242 (1997).
7. De Matteis, M. A. & Luini, A. Exiting the Golgi complex. *Nature Rev. Mol. Cell Biol.* **9,** 273–284 (2008).
8. Shorter, J. & Warren, G. Golgi architecture and inheritance. *Annu. Rev. Cell Dev. Biol.* **18,** 379–420 (2002).
9. Short, B., Haas, A. & Barr, F. A. Golgins and GTPases, giving identity and structure to the Golgi apparatus. *Biochim. Biophys. Acta* **1744,** 383–395 (2005).
10. Altan-Bonnet, N. *et al.* Golgi inheritance in mammalian cells is mediated through endoplasmic reticulum export activities. *Mol. Biol. Cell* **17,** 990–1005 (2006).
11. Diao, A. *et al.* The coiled-coil membrane protein golgin-84 is a novel rab effector required for Golgi ribbon formation. *J. Cell Biol.* **160,** 201–212 (2003).
12. Lowe, M. *et al.* Cdc2 kinase directly phosphorylates the cis-Golgi matrix protein GM130 and is required for Golgi fragmentation in mitosis. *Cell* **94,** 783–793 (1998).
13. Wang, Y. *et al.* A direct role for GRASP65 as a mitotically regulated Golgi stacking factor. *EMBO J.* **22,** 3279–3290 (2003).
14. Chiu, R., Novikov, L., Mukherjee, S. & Shields, D. A caspase cleavage fragment of p115 induces fragmentation of the Golgi apparatus and apoptosis. *J. Cell Biol.* **159,** 637–648 (2002).
15. Lowe, M., Lane, J. D., Woodman, P. G. & Allan, V. J. Caspase-mediated cleavage of syntaxin 5 and giantin accompanies inhibition of secretory traffic during apoptosis. *J. Cell Sci.* **117,** 1139–1150 (2004).
16. Mancini, M. *et al.* Caspase-2 is localized at the Golgi complex and cleaves golgin-160 during apoptosis. *J. Cell Biol.* **149,** 603–612 (2000).
17. Rajalingam, K. *et al.* Epithelial cells infected with *Chlamydophila pneumoniae* (*Chlamydia pneumoniae*) are resistant to apoptosis. *Infect. Immun.* **69,** 7880–7888 (2001).
18. Fan, T. *et al.* Inhibition of apoptosis in chlamydia-infected cells: blockade of mitochondrial cytochrome c release and caspase activation. *J. Exp. Med.* **187,** 487–496 (1998).
19. Hood, J. L., Brooks, W. H. & Roszman, T. L. Differential compartmentalization of the calpain/calpastatin network with the endoplasmic reticulum and Golgi apparatus. *J. Biol. Chem.* **279,** 43126–43135 (2004).
20. Goll, D. E. *et al.* The calpain system. *Physiol. Rev.* **83,** 731–801 (2003).
21. Suzuki, K. & Sorimachi, H. A novel aspect of calpain activation. *FEBS Lett.* **433,** 1–4 (1998).
22. Puthenveedu, M. A. *et al.* GM130 and GRASP65-dependent lateral cisternal fusion allows uniform Golgi-enzyme distribution. *Nature Cell Biol.* **8,** 238–248 (2006).

23. Holthuis, J. C. & Levine, T. P. Lipid traffic: floppy drives and a superhighway. *Nature Rev. Mol. Cell Biol.* **6,** 209–220 (2005).
24. Marie, M., Sannerud, R., Avsnes Dale, H. & Saraste, J. Membrane traffic in the secretory pathway: Take the 'A' train: on fast tracks to the cell surface. *Cell. Mol. Life Sci.* **65,** 2859–2874 (2008).
25. Machuy, N. *et al.* A global approach combining proteome analysis and phenotypic screening with RNA interference yields novel apoptosis regulators. *Mol. Cell. Proteomics* **4,** 44–55 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors would like to thank A. Greiser, C. Goosmann, B. Laube, M. Wicht and E. Ziska for technical support; M. A. De Matteis for the gift of the GFP–GM130 fusion plasmid; H. P. Hauri for the provision of the anti-GPP130 antibody and helpful discussions; and K. Astrahantseff, S. and J. Heuer, T. Wolff and L. Ogilvie for critically reading the manuscript and helpful suggestions. This work was financially supported by the Senate of Berlin and the BMBF through the RiNA Network.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.F.M. (meyer@mpiib-berlin.mpg.de).

METHODS

Reagents. Antibodies were obtained from the following sources: mouse anti-tubulin antibody (Sigma-Aldrich), mouse anti-golgin-84 (raised against the C-terminus), mouse anti-giantin, mouse anti-GM130, mouse anti-p230 (BD Biosciences), mouse anti-GPP130 (gift from H.-P. Hauri), mouse anti-*Chlamydia* HSP60 (Axxora), rabbit anti-LPS (Milan Analytica). WEHD-FMK, DEVD-FMK and caspase inhibitor IV were purchased from R&D Systems, Calbiochem or Merck Biosciences. BODIPY FL C5-ceramide in complex with BSA and GS-II Alexa Fluor 594 was purchased from Invitrogen, and Hoechst 33342 was from Sigma-Aldrich.

Live-cell imaging of infected EGFP–GM130-expressing HeLa cells. Cells were seeded into glass-bottom-like dishes (Ibidi) and transfected with a plasmid expressing human GM130 fused to EGFP (gift from M. A. De Matteis) using Lipofectamine 2000 (Invitrogen) according to the manufacturer's guidelines. The cells were subsequently either infected or mock-infected, and recorded 26 h after infection at 37 °C on a Zeiss Axiovert 200M microscope with a Plan-Neofluar 100×/1.3 objective (Jena) over a period of 20 h. Every 3 min, a set of two images (phase contrast, GFP) was taken with a Hamamatsu Orca ER camera. The system was controlled using Openlab software (Improvision), and individual frame overlays and videos were prepared using Volocity software (Improvision).

Ultrastructural analysis of cells by electron microscopy. Normal HeLa cells, golgin-84 knockdown cells and control knockdown cells were infected with *C. trachomatis* at a MOI of 2. Cells were fixed 24 h after infection with 2.5% glutaraldehyde, post-fixed with 1% OsO_4 for 45 min and contrasted with tannic acid and uranyl acetate. Specimens were dehydrated in a graduated ethanol series and embedded in PolyBed (Polysciences Europe GmbH). After polymerization, blocks were cut at 60–80 nm, contrasted with lead citrate and analysed in a LEO 906E TEM (Zeiss SMT) equipped with a Morada camera (SIS).

Treatment of cells with caspase inhibitors. Caspase inhibitor IV (40 μM), Z-WEHD-FMK (80 μM), Z-DEVD-FMK (80 μM), calpeptin (25 μg ml^{-1}), calpain inhibitor III (100 μM), MG-132 (2.5 μM), PS1 (2.5 μM) and CA-074 Me (50 μM) was added to infected or mock-infected HeLa cells 9 h after infection (caspase inhibitors) or 8 h after infection (calpain inhibitors). Cells were analysed 24–28 h after infection using immunofluorescence and immunoblotting. To test effects on *C. trachomatis* propagation, Z-WEHD-FMK was added 9 h after infection to infected HeLa cells that were transfected with siRNA targeting luciferase or *golgin-84*. At 48 h after infection newly formed *C. trachomatis* bacteria were titrated on fresh HeLa cells.

siRNA transfection. HeLa cells were seeded into 12-well plates, grown to 50–70% confluency, and transfected using RNAiFect (Qiagen) according to the manufacturer's guidelines. In brief, 1 μg of specific siRNA was added to EC-R buffer and incubated with 6 μl RNAiFect transfection reagent in a total volume of 100 μl. After 10–15 min the liposome/RNA mixture was added to the cells with 600 μl cell culture medium. After 1 day, cells were trypsinized and seeded into new cell culture plates depending on the experiments. Three days after transfection, the cells were infected and incubated as indicated above.

Determination of cell viability by WST-1 assay. The viability of knockdown cells was determined using the WST-1 reagent (Roche). The assay was performed in triplicate using a 96-well format three days after transfection. The WST-1 reagent was diluted 1:5 in cell culture medium containing 5% FBS, added directly to the cultures and incubated for 4 h at 37 °C and 5% CO_2. As a negative control, untreated cells were lysed by Triton X-100 before addition of the reagent. The optical density at 450 nm was measured and absorption of untreated cells was set to 100%.

Immunofluorescence and microscopy. Cells were seeded onto cover slips to visualize the Golgi apparatus in either infected or siRNA-transfected HeLa cells. At the indicated time points, cells were fixed with 2% PFA for 30 min at room temperature. Cells were then permeabilized with 0.2% Triton X-100/0.2% BSA in PBS for 30 min. Different Golgi markers were detected using specific antibodies, and bacteria were detected either using rabbit anti-LPS or mouse anti-HSP60 antibodies followed by specific fluorescently labelled secondary antibodies and mounted in MOWIOL. Images were taken with a Leica TCS-SP confocal microscope and processed using Adobe Photoshop 6.0.

Immunoblotting. Infected or siRNA-transfected HeLa cells were lysed in RIPA buffer at the indicated time points. Protein concentrations were determined using the Pierce BCA kit (Perbio Science), and 20 μg of total protein were analysed per lane on a reducing SDS-polyacrylamide gel. After separation, proteins were transferred onto a PVDF membrane by tank blotting. Specific antibodies were incubated with the membrane to detect antigens, followed by visualization of the appropriate secondary antibodies using ECL reagent, as described previously[26].

Quantification of Golgi fragments. Confocal images of specific samples were used to quantify Golgi fragmentation. The numbers of Golgi elements in 25 cells per experiment were counted after applying a fixed threshold to all images using the Analyse Particles function in ImageJ software. Three independent experiments were performed in duplicate to analyse fragmentation: on infection, infection plus Z-WEHD-FMK treatment and staurosporine treatment.

Treatment with staurosporine. To induce Golgi fragmentation by staurosporine, HeLa cells were incubated for 4 h at 35 °C with 2 μM staurosporine, and subsequently immunostained with the monoclonal mouse anti-GM130 antibody.

siRNA. All siRNAs were designed and purchased from Qiagen, and validated at the Max Planck Institute for Infection Biology for their ability to knock down mRNA expression of target genes by more than 70% in comparison with control cells transfected with siRNA for luciferase.

Validation of RNAi by quantitative PCR. siRNA validation was performed according to ref. 25. Briefly, one day before transfection 3,000 cells per well were seeded onto a 96-well plate. Transfection was performed with a final siRNA concentration of 56 nM with 0.25 μl RNAiFect (Qiagen) using luciferase targeting siRNA (target sequence in XYN_{19} format: AACUUACGCUGAGUACUUCGA) as a control and the following target-specific siRNAs: *golgin-84* (CTGAGTTTAGTGGTCCTAATA), *GM130* (CAGGCTGGAGTTATACAAGAA), *golgin-245* (CAGGAAATACATGAAATCCAA), *giantin* (AACTTCATGCGAAGGCCAAAT), *GPP130* (CAGGAGGACAATGTTGATGAA), *CASP* (CAGCGCCTGCACGATATTGAA). Knockdown measurements were performed independently three times. After 48 h, RNA was isolated using the RNeasy 96 BioRobot 8000 system (Qiagen). The relative amount of target mRNA was determined by quantitative PCR using the Quantitect SYBR Green RT–PCR kit following the manufacturer's instructions (Qiagen) and the following primers: *GAPDH* forward 5′-GGTATCGTGGAAGGACTCATGAC-3′, *GAPDH* reverse 5′-ATGCCAGTGAGCTTCCCGTTCAG-3′, *golgin-84* forward 5′-AATGCACCACGACCAACCA-3′, *golgin-84* reverse 5′-AGGCAATTGGCCTTCTTGC-3′, *GM130* forward 5′-AATATCAGCAGAGGAATAGCCCT-3′, *GM130* reverse 5′-CAGCATTGTCCTTGGGTGTAT-3′, *golgin-245* forward 5′-ATGTATATGCAACAACTGTGGGG-3′, *golgin-245* reverse 5′-CGAGGTGAAGTAAACATCAGCC-3′, *giantin* forward 5′-CCCTAGACCCTGAATTACACCAA-3′, *giantin* reverse 5′-GGCAGAACAGTCCCTCCTTG-3′, *GPP130* forward 5′-CCCTCTCCGCCCAGTTACA-3′, *GPP130* reverse 5′-CTCCTCGTGTTGGCTTTTCA-3′, *CASP* forward 5′-AAAGACCAGCCTGAAAGTCGG-3′, *CASP* reverse 5′-CCAGGGATGAGCTGAAAAAGT-3′. The relative expression levels of target mRNA were normalized against control transfected cells. *GAPDH* was used as an internal standard.

Generation of stable golgin-84 knockdown cell lines. The golgin-84 shRNA construct oligonucleotides (Metabion) were annealed and ligated into the lentiviral vector pLVTHM targeting the following sequence in the 3′ UTR of *golgin-84*: GAGAACAGUGCACAAGAUUAU. Cells transduced with lentiviruses coding for a firefly luciferase shRNA (target sequence AACUUACGCUGAGUACUUCGA) were used as a control. All constructs were verified by sequencing. Viruses carrying the shRNAs were produced by transfecting 293T cells with pLVTHM golgin84-15 or pLVTHM luciferase together with viral packaging vectors (psPAX2, pMD2G) by calcium phosphate transfection. Viruses were harvested from the supernatant 48 h after transfection, filtrated through a 0.45 μm filter and applied to HeLa cells for lentiviral infection in the presence of polybrene (5 μg ml^{-1}). GFP-positive cells were selected 6 days after infection, and single cells were transferred to a 96-well plate using FACS sorting. All vectors were obtained from D. Trono.

Generation of stable golgin-84-expressing cell lines. The N-terminal truncated golgin-84 mutants were amplified from pENTR221-golga5 obtained from RZPD by Expand HighFidelity Taq (Roche) using specific primer pairs (Supplementary Table 1). All primers were from MWG. Products were cloned into pDONR221 (Invitrogen) by clonase II (Invitrogen) reaction and further subcloned into pLentiV5/DEST (Invitrogen). These vectors were used to generate specific lentiviruses, as described above. Golgin-84 knockdown cells were infected with the respective virus and positive cells were selected using 10 μg ml^{-1} blasticidin (Merck Biosciences). Generated cell lines were passaged in the presence of 2 μg ml^{-1} blasticidin. For infection experiments with *C. trachomatis*, blasticidin was removed one day before infection. Plasmids and corresponding cell lines are listed in Supplementary Information 2.

Lectin binding. Binding of GS-II Alexa Fluor 594 has been described previously[22]. Briefly, 28 h after infection *C. trachomatis*-infected or uninfected cells grown on cover slips were washed with ice-cold PBS^{++} (containing $MgCl_2$ and $CaCl_2$). 100 μg ml^{-1} fluorescent GS-II in PBS^{++} was added onto the cells and incubated for 2 h at 4 °C. The cells were washed with PBS and then fixed with 2% PFA for 30 min at room temperature. Host cell and *Chlamydia* DNA was visualized by Hoechst staining. Samples were analysed using an inverted confocal microscope (SP5, Leica).

Identification of golgin-84 cleavage site by MS. Full-length human golgin-84 was amplified from pENTR221-golga5 by Expand HighFidelity Taq (Roche) using specific primer pairs that contained a C-terminal Myc tag 5′-AAAAA-GCAGGCTGAACCATGTCTTGGTTTGTTGATCTTGC-3′ and 5′-AGAAAG-CTGGGTATCACTACAGATCTTCTTCAGAAATAAGTTTTTGTTCTTTGCC-ATATGGTTGGTCGTGGTGC-3′. Products were cloned into pDONR221 (Invitrogen) by clonase II (Invitrogen) reaction and further subcloned into pDest760. Tagged golgin-84 was transiently expressed and cells were infected with *C. trachomatis*. Thirty-six hours after infection cells were lysed in RIPA buffer containing protease inhibitor cocktail (Roche) and tagged golgin-84 was immunoprecipitated by anti-Myc antibodies 9E10 (Santa Cruz) followed by Dynabeads Protein G (Invitrogen). Precipitated golgin-84 was subjected to SDS–PAGE and the gel was stained with Serva DensiStain Blue G (Serva, Electrophoresis GmbH). The 65-kDa golgin-84 cleavage product was cut from the gel and digested with modified trypsin (Promega) according to the supplier's instructions. Eluted peptides were analysed by Nano-LC-MS/MS (Nano-Acquity, Waters, combined with a LTQ-Orbitrap, Thermo Fisher). A linear gradient from water to acetonitrile (both with 0.1% formic acid) was used to elute the peptides with a flow of 200 nl min^{-1} from a self-prepared Nano-RP-column (Reprosil-Pur 300 C18, 3 μm, Dr. Maisch, packed in PicoTip Emitter, 75 μm × 150 mm, New Objective). Throughout the whole run, MS and MS/MS spectra were collected in data-dependent acquisition mode. For protein identification, an InHouse version of MASCOT-server (MatrixScience) was used to search against the Swiss-Prot database (Release 55.4).

26. Heuer, D., Brinkmann, V., Meyer, T. F. & Szczepek, A. J. Expression and translocation of chlamydial protease during acute and persistent infection of the epithelial HEp-2 cells with *Chlamydophila* (*Chlamydia*) *pneumoniae*. *Cell. Microbiol.* **5,** 315–322 (2003).

LETTERS

Messenger RNA targeting to endoplasmic reticulum stress signalling sites

Tomás Aragón[1,2]*, Eelco van Anken[1,2]*, David Pincus[1,2], Iana M. Serafimova[1,2], Alexei V. Korennykh[1,2], Claudia A. Rubio[1,2] & Peter Walter[1,2]

Deficiencies in the protein-folding capacity of the endoplasmic reticulum (ER) in all eukaryotic cells lead to ER stress and trigger the unfolded protein response (UPR)[1–3]. ER stress is sensed by Ire1, a transmembrane kinase/endoribonuclease, which initiates the non-conventional splicing of the messenger RNA encoding a key transcription activator, Hac1 in yeast or XBP1 in metazoans. In the absence of ER stress, ribosomes are stalled on unspliced *HAC1* mRNA. The translational control is imposed by a base-pairing interaction between the *HAC1* intron and the *HAC1* 5′ untranslated region[4]. After excision of the intron, transfer RNA ligase joins the severed exons[5,6], lifting the translational block and allowing synthesis of Hac1 from the spliced *HAC1* mRNA to ensue[4]. Hac1 in turn drives the UPR gene expression program comprising 7–8% of the yeast genome[7] to counteract ER stress. Here we show that, on activation, Ire1 molecules cluster in the ER membrane into discrete foci of higher-order oligomers, to which unspliced *HAC1* mRNA is recruited by means of a conserved bipartite targeting element contained in the 3′ untranslated region. Disruption of either Ire1 clustering or *HAC1* mRNA recruitment impairs UPR signalling. The *HAC1* 3′ untranslated region element is sufficient to target other mRNAs to Ire1 foci, as long as their translation is repressed. Translational repression afforded by the intron fulfils this requirement for *HAC1* mRNA. Recruitment of mRNA to signalling centres provides a new paradigm for the control of eukaryotic gene expression.

In vitro studies indicate that the information required for *HAC1* mRNA splicing is confined to the intron and the regions surrounding the splice junctions[8]. Surprisingly, *in vivo* splicing of *HAC1* mRNA was greatly diminished when its 3′ untranslated region (3′ UTR) was replaced by the 3′ UTRs of other yeast mRNAs, such as that of actin (*ACT1*, Fig. 1b) or 3-phosphoglycerate kinase (*PGK1*, data not shown). Consistent with this finding, cells bearing a chimaeric *HAC1* gene with the 3′ UTR of *ACT1*, *HAC1*-3′*act1*, expressed Hac1 protein at trace levels that were too low to mount a functional UPR and failed to grow in ER stress conditions (Fig. 1b). Thus, the *HAC1* 3′ UTR harbours an element important for *HAC1* mRNA splicing *in vivo*.

Mutational probing experiments (not shown) indicate that the *HAC1* 3′ UTR contains a prominent, extended stem-loop (Fig. 1c). Interestingly, two short sequence motifs within the stem-loop are highly conserved among all *HAC1* orthologues identified; eight representatives are shown in Fig. 1d. The sequence motifs map to opposite strands and are juxtaposed in the distal part of the stem, constituting a 3′ UTR bipartite element (3′ BE; Fig. 1c, 3′ BE in red).

To assess the importance of the 3′ BE for *HAC1* mRNA splicing *in vivo*, we used a splicing reporter in which we replaced the first 648 nucleotides of the *HAC1* coding sequence in the first exon with that of green fluorescent protein (GFP; Fig. 1a, green bar). This reporter allowed us to monitor the effect of 3′ UTR mutations on mRNA splicing in cells that can mount a functional UPR, sustained by endogenous *HAC1* mRNA. The splicing reporter mRNA was efficiently spliced on UPR induction. In contrast, splicing was greatly diminished when the 3′ BE was deleted (Δ3′ BE, Fig. 1e). Consistent with these results, deletion of the 3′ BE in *HAC1* severely reduced *HAC1* mRNA splicing and impaired cell survival under ER stress conditions (Fig. 1e). Only residual splicing of endogenous *HAC1* mRNA occurred in the absence of the 3′ BE, indicating that the 3′ BE accounts in large part for the contribution of the 3′ UTR to *HAC1* mRNA splicing. Insertion of a 64-nucleotide 3′ UTR fragment containing the central portion of the stem including the 3′ BE (Fig. 1d, enlarged on right) into the splicing reporter bearing the *ACT1* 3′ UTR restored splicing greatly (Fig. 1f).

To test whether the 3′ UTR affects the ability of Ire1 endonuclease to bind or catalyse the cleavage of *HAC1* mRNA, we reconstituted the intron excision reaction *in vitro*. Ire1 cleaved *HAC1* mRNA with the same rate in the presence or absence of the 3′ BE (Fig. 1g). Thus, the 3′ BE is not required for splicing *in vitro*.

The importance of the 3′ BE for *HAC1* mRNA splicing *in vivo* indicated that it may serve to target the mRNA to sites in the cell where splicing takes place. To test this notion, we visualized Ire1 protein and *HAC1* mRNA *in vivo*, using the imaging constructs depicted in Fig. 2a. For Ire1, we inserted a GFP or mCherry into the cytosolic portion of Ire1 adjacent to its transmembrane region. For *HAC1* mRNA, we inserted 16 copies of a U1A binding site into the 3′ UTR downstream of the 3′ BE. The mRNA can then be visualized by co-expression of a GFP-tagged U1A-RNA-binding protein that docks to the U1A binding sites[9]. Both Ire1 and *HAC1* mRNA imaging constructs fully restored growth of *ire1*Δ and *hac1*Δ (Fig. 2b) cells under ER stress. In the absence of stress, Ire1–GFP co-localized with the ER marked by Sec63–mCherry (Fig. 2c). Most *HAC1*^{U1A} mRNA displayed a grainy signal dispersed throughout the cytosol (Fig. 2d), with a fraction of *HAC1* mRNA signal also found at the ER in agreement with previous observations[10].

Induction of ER stress notably altered the localization of both Ire1 and *HAC1* mRNA. Most Ire1 (82 ± 6%; see Methods) clustered into distinct foci localized both to the nuclear envelope and to the cortical ER (Fig. 2c–e), in agreement with recent observations[11]. *HAC1* mRNA strongly co-localized (co-localization index (CI) of 56 ± 10; see Methods, Fig. 2e) with Ire1 in foci (Fig. 2d, arrowheads). This recruitment is specific, because control *PGK1*^{U1A} mRNA remained dispersed in the cytosol under ER stress conditions (Fig. 2f).

Clustering of mRNAs in cytosolic foci is not unprecedented. Several stresses, such as nutrient starvation, cause aggregation of untranslated mRNAs into processing bodies (P-bodies) where they are stored and/or degraded[12]. The Ire1/*HAC1* mRNA clusters, however, are distinct

[1]Department of Biochemistry and Biophysics, and [2]Howard Hughes Medical Institute, University of California at San Francisco, San Francisco, California 94158-2517, USA.
*These authors contributed equally to this work.

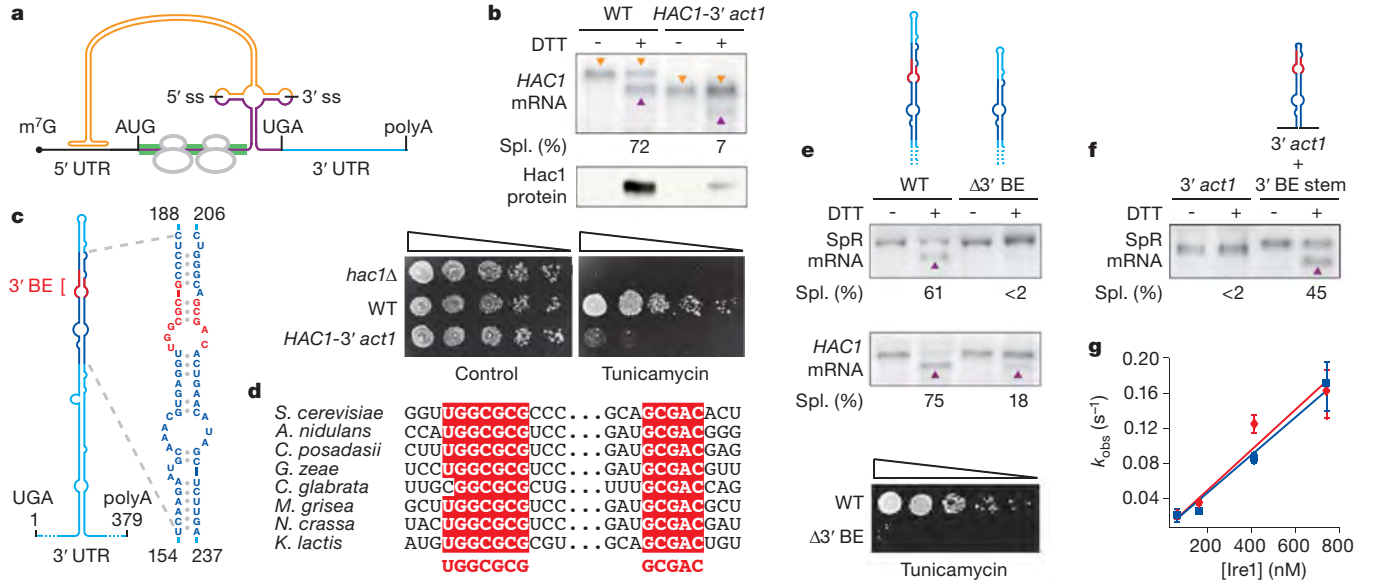


Figure 1 | A conserved element in the 3′ UTR of *HAC1* mRNA is required for splicing *in vivo*, but not *in vitro*. **a**, Schematic of *HAC1* mRNA. The *HAC1* open reading frame (ORF) is divided into two exons (purple). The intron (orange) base pairs with the 5′ UTR (black), causing stalling of ribosomes (grey). Ire1 cleaves the intron at the indicated splice sites (5′ ss and 3′ ss). The green bar depicts where the GFP ORF replaces the *HAC1* sequence in the splicing reporter. The 3′ UTR is indicated in light blue. The 5′ cap (m^7G), start codon (AUG), stop codon (UGA) and polyadenylation signal (polyA) are indicated. **b**, **e**, **f**, Northern blot of *HAC1* or splicing reporter (SpR) mRNA variants before or after ER stress induction with DTT (10 mM) for 45 min. Purple triangles denote spliced mRNAs; orange triangles denote unspliced mRNAs (only in **b**). Percentage mRNA splicing (Spl. (%)) is indicated. Yeast strains harbour: a genomic *HAC1* copy with its own (WT) or *ACT1*'s 3′ UTR sequence (*HAC1-3′act1*; **b**, top); a genomic copy of SpR (**e**, top) or *HAC1* (**e**, middle) bearing either the wild-type (WT) or the Δ3′ BE mutant 3′ UTR of *HAC1*, as depicted; or a genomic copy of SpR with the 3′ UTR of *ACT1* with (3′*act1* + 3′ BE stem) or without (3′*act1*) an insertion of the 64-nucleotide element (shown in expanded view in **c**), as depicted (**f**). **b**, Middle: western blot of haemagglutinin (HA)-tagged Hac1 protein from lysates from strains as in the top panel of **b**. **b**, **e**, Viability assay by 1:5 serial dilutions of *hac1*Δ or strains as in the top panel of **b** or the middle panel of **e**, spotted onto solid media with or without 0.2 μg ml^{-1} of the ER-stress-inducer tunicamycin. Plates were photographed after 3 days growing at 30 °C. **c**, Schematic of the *HAC1* 3′ UTR stem-loop structure with the 3′ BE (red) in a region (dark blue) that is shown in expanded view to the right; positional numbering is from UGA stop codon. **d**, Alignment of the 3′ BE in *HAC1* homologues (*Saccharomyces cerevisiae*, *Aspergillus nidulans*, *Coccidioides posadasii*, *Gibberella zeae*, *Candida glabrata*, *Magnaporthea grisea*, *Neurospora crassa*, *Kluyveromyces lactis*). **g**, An *in vitro* intron excision reaction was performed as described[8] with Ire1 concentrations (50 nM, 150 nM, 400 nM and 730 nM) of wild-type (red diamonds) or Δ3′ BE (blue squares) *HAC1* mRNA as substrates. Error bars show standard errors of single-exponential fitting.

from P-bodies: on glucose depletion *HAC1* mRNA did cluster into P-bodies marked by Lsm1–mCherry[13]. In contrast, under ER stress conditions Lsm1–mCherry did not co-localize with *HAC1* mRNA foci but remained dispersed throughout the cytosol (Fig. 2g). Thus, the Ire1/*HAC1* mRNA foci constitute previously unknown sites of mRNA clustering in the cytosol that are specific for the UPR.

We next determined the role of each of the three key UPR players—Ire1, *HAC1* mRNA and tRNA ligase—in organizing the foci. In *rlg1-100* cells bearing mutant tRNA ligase defective in UPR signalling[5], co-clustering of Ire1 and *HAC1* mRNA occurred normally (Fig. 2h). This result is consistent with the fact that cleavage of *HAC1* mRNA by Ire1 is not dependent on the subsequent ligation step[5]. Likewise, *HAC1* mRNA was not required for Ire1 clustering, because Ire1–GFP formed foci in *hac1*Δ cells (Fig. 2h and ref. 11). Conversely, *HAC1* mRNA failed to form foci in *ire1*Δ cells (Fig. 2h). Thus, clustering of Ire1 in response to ER stress is epistatic to *HAC1* mRNA clustering.

Having established that *HAC1* mRNA is targeted to Ire1 foci in an ER-stress-driven manner, we assessed the role of the 3′ UTR of *HAC1* mRNA in the process. To this end, we added the U1A visualization module to the splicing reporter used in Fig. 1e (SpR^{U1A}). The SpR^{U1A} mRNA containing a wild-type *HAC1* 3′ UTR co-localized with Ire1–mCherry in foci (CI: 64 ± 20; Fig. 2i). In contrast, co-localization with Ire1 foci of the SpR^{U1A} mRNA lacking the 3′ BE was minimal (CI: 4 ± 6; Fig. 2i), at levels comparable to the control $PGK1^{U1A}$ mRNA (CI, 3 ± 4). Thus, the stem-loop structure in the 3′ UTR of *HAC1* mRNA—with the 3′ BE at its core—indeed serves as a targeting element that guides *HAC1* mRNA to Ire1 foci to allow splicing *in vivo* and cell survival under ER stress.

We next followed a time course of foci formation and downstream signalling on induction of ER stress. Clustering of Ire1 into foci and recruitment of *HAC1* mRNA (Fig. 3a, b) or of SpR^{U1A} mRNA (Supplementary Fig. 1) into these foci correlated well with the onset of *HAC1* mRNA splicing and Hac1 protein production (Fig. 3c). These findings show that Ire1 and *HAC1* mRNA clustering is geared to transduce ER stress rapidly. Under conditions in which ER stress builds up more gradually, the encounter of Ire1 and $HAC1^{U1A}$ mRNA in foci likewise paralleled the signalling response, but at a slower pace (Supplementary Fig. 2). The synchronicity of Ire1/*HAC1* mRNA clustering and downstream signalling events underscores that the foci constitute functional mRNA-splicing centres.

Ire1 clusters in only ~3–10 foci per cell. Because yeast contains ~200–300 molecules of Ire1 per cell[14], the foci are composed of a few tens of Ire1 molecules each, indicating that the foci harbour higher-order oligomers of Ire1. From the crystal structure of the Ire1 ER-luminal domain, we identified two separate dimerization interfaces, both of which are essential for optimal UPR signalling, indicating that oligomerization is important for cells to mount a robust UPR[15] (Fig. 3d). Accordingly, simultaneous disruption of both interfaces notably reduced *HAC1* mRNA splicing and cell growth under ER stress conditions, whereas the single-interface disruptions, which still can form Ire1 dimers by means of one interface, displayed intermediate splicing and growth phenotypes (Fig. 3e). Disruption of either interface prevented foci formation (Fig. 3f, Supplementary Fig. 3 and

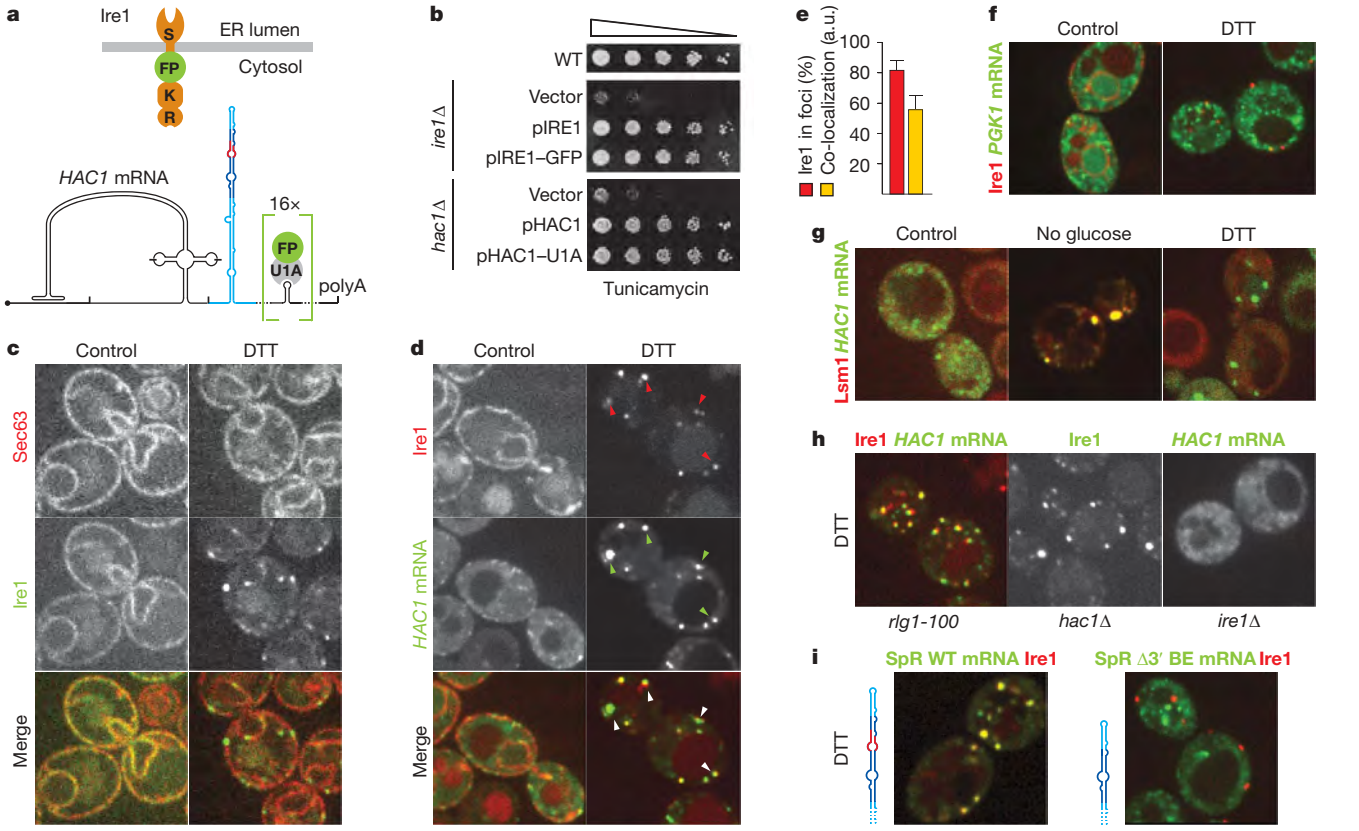


Figure 2 | In response to ER stress *HAC1* mRNA localizes to Ire1 foci in a 3′ BE-dependent manner. **a**, Schematic of Ire1 and *HAC1* mRNA imaging constructs: Ire1 has an ER-luminal stress-sensing domain (S), and has a kinase (K) and an endoribonuclease domain (R) at its cytosolic face. GFP or mCherry (FP) was inserted between the transmembrane region and the kinase domain. 16 U1A binding sites were inserted into the 3′ UTR of *HAC1* mRNA downstream of the stem-loop. Binding of GFP-tagged U1A protein allows visualisation of the mRNA. **b**, Viability assay under ER stress conditions (0.2 μg ml^{-1} tunicamycin) of wild-type (WT) or *ire1*Δ yeast complemented with empty vector or with centromeric plasmids bearing a wild-type (pIRE1) or the GFP-tagged imaging copy of Ire1 (pIRE1–GFP; top), or of *hac1*Δ yeast complemented with either empty plasmid or with a 2-μm plasmid bearing a wild-type (pHAC1) or the U1A-tagged imaging copy of *HAC1* (pHAC1–U1A; bottom). **c**, **d**, Localization of Sec63–mCherry and Ire1–GFP (**c**) or Ire1–mCherry and *HAC1*^{U1A} mRNA decorated with U1A–GFP (**d**) before (left panels, control) and after (right panels, DTT) induction of ER stress. Arrowheads in **d** denote Ire1/*HAC1* mRNA foci. **e**, Histogram depicting the percentage of Ire1 signal in foci (red bar) and the co-localization index for *HAC1*^{U1A} mRNA recruitment into Ire1 foci expressed in arbitrary units (yellow bar); means and s.e.m. are shown, $n = 9$. **f**, Localization of Ire1–mCherry and *PGK1*^{U1A} mRNA under normal (left panel, control) and ER stress (right panel, DTT) conditions. **g**, Localization of Lsm1–mCherry and *HAC1*^{U1A} mRNA without stress (left panel, control), after nutrient starvation for 10 min (middle panel, no glucose), or after induction of ER stress (right panel, DTT). **h**, **i**, Localization of Ire1–mCherry (red font), Ire1–GFP (green font), *HAC1*^{U1A}, or splicing reporter with 16 U1A hairpins as in *HAC1*^{U1A} (SpR^{U1A}) either with or without the Δ3′ BE deletion after induction of ER stress (DTT). **c**–**i**, ER stress was induced with 10 mM DTT for 45 min; imaging was performed in *ire1*Δ cells, complemented with Ire1 imaging constructs, except in **h** in which the cells were *hac1*Δ or *rlg1-100*, where indicated.

ref. 11), indicating that Ire1 oligomerization is the organizing principle for UPR signalling foci. Importantly, the inability of Ire1 to form foci impaired *HAC1*^{U1A} mRNA recruitment (Fig. 3f and Supplementary Fig. 3). Thus, when Ire1 fails to oligomerize, *HAC1* mRNA recruitment becomes rate limiting. In agreement, we found that artificially induced dimerization[16] of Ire1 supported *HAC1* mRNA splicing and cell survival under ER stress conditions only to the level of the single-interface mutants and did not support Ire1 foci formation (Supplementary Fig. 4). We conclude that robust Ire1 oligomerization and *HAC1* mRNA targeting serve to concentrate both key UPR components into foci to ensure efficient RNA processing and ER stress signalling.

HAC1 mRNA is no longer a substrate for Ire1 after removal of its intron, indicating that the spliced *HAC1* mRNA should disengage from Ire1 foci and not be recruited again. Accordingly, SpR^{U1A} mRNA lacking the intron displayed reduced targeting to foci (CI: 17 ± 9) compared to wild-type SpR^{U1A} mRNA (Fig. 4a, b), although targeting was not as markedly reduced as when the 3′ BE was deleted (Figs 2i and 4b). In further support, overexpression of SpR^{U1A} mRNA containing the intron reduced splicing of endogenous *HAC1* mRNA, presumably by competitively saturating Ire1 after being targeted there, but did not do so when SpR^{U1A} mRNA lacked either the 3′ BE or the intron (Fig. 4c). These observations indicate that the 3′ BE alone is not sufficient for efficient targeting. In agreement, insertion of the 3′ UTR stem of *HAC1* (Fig. 1c) into the 3′ UTR of *PGK1* could not facilitate recruitment of this heterologous mRNA to Ire1 foci (Fig. 4d). Thus, the intron and 3′ BE cooperate to effect *HAC1* mRNA targeting.

The intron keeps *HAC1* mRNA translationally silent (Fig. 1a), indicating that translational repression may be key to *HAC1* targeting similar to the situation in other mRNA targeting mechanisms, as observed for *ASH1* mRNA[17]. To test this hypothesis, we inserted a small stem-loop into the 5′ UTR of the *PGK1*^{U1A} mRNA to repress its translation (ref. 17 and Fig. 4e). When we expressed *PGK1*^{U1A} mRNA containing both the small stem-loop in the 5′ UTR and the *HAC1* 3′ BE-containing stem in the 3′ UTR, we found that this mRNA efficiently targeted to Ire1 foci (CI: 60 ± 19, Fig. 4f, h). Conversely, the corresponding mRNA lacking the 3′ BE was not targeted (Fig. 4g, h). We conclude that the 3′ BE-containing stem is both necessary and sufficient to target a heterologous mRNA to UPR-induced Ire1 foci, provided that its translation is on hold. Translational repression,

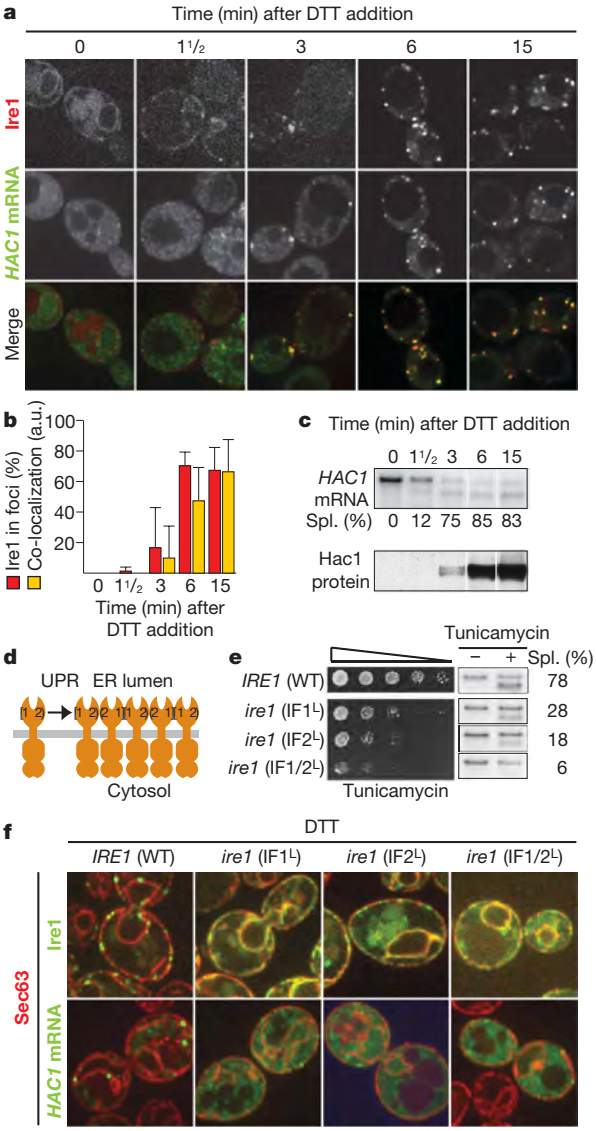


Figure 3 | The *HAC1* mRNA/Ire1 foci are functional UPR signalling centres. **a**, Localization of Ire1–mCherry and *HAC1*^{U1A} mRNA decorated with U1A–GFP. **b**, Quantification of the percentage of Ire1 signal in foci (red bars) and of the co-localization index for *HAC1*^{U1A} mRNA recruitment into Ire1 foci expressed in arbitrary units (yellow bars; means and s.e.m., $n = 5$). **c**, Northern blot of *HAC1* mRNA (top) and western blot of Hac1 protein (bottom). **a**–**c**, Samples were taken at indicated times after induction of ER stress with 10 mM DTT. **d**, Schematic of Ire1 oligomerization via interfaces 1 and 2. **e**, Viability assay under ER stress conditions (0.2 μg ml^{-1} tunicamycin) and northern blot of *HAC1* mRNA collected from *ire1*Δ yeast complemented with wild-type *IRE1* or of *ire1* mutants that are defective in dimerization at luminal interface 1 (IF1^{L}), 2 (IF2^{L}) or both (IF1/2^{L}) before or after treatment with 1 μg ml^{-1} tunicamycin for 1 h. **f**, Localization of Sec63–mCherry, Ire1–GFP and *HAC1*^{U1A} mRNA. Imaging was performed in *ire1*Δ yeast complemented with wild-type or 'IF' mutants, either GFP-tagged (top) or untagged (bottom). ER stress was induced with 10 mM DTT for 45 min. Separate channels are displayed in Supplementary Fig. 3.

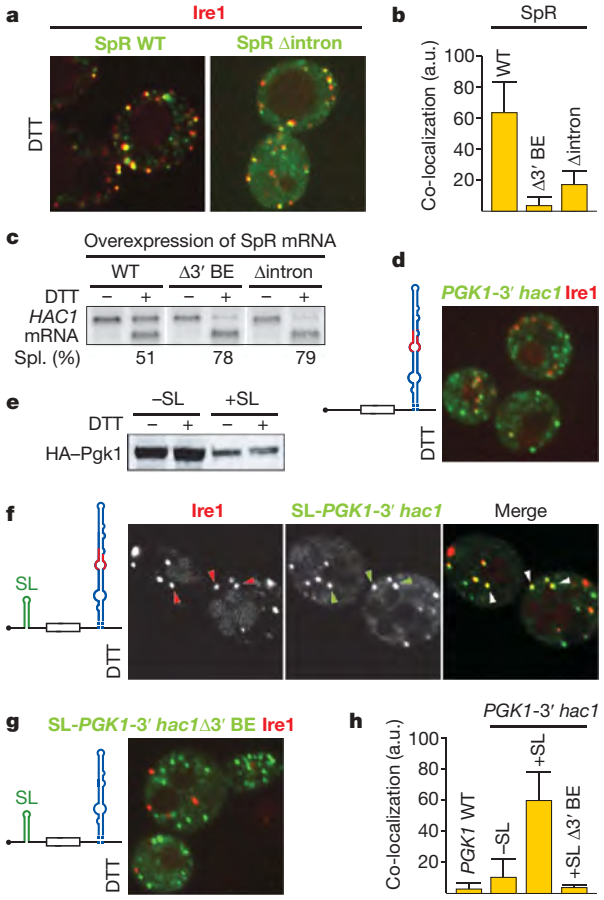


Figure 4 | Translational repression is a prerequisite for mRNA targeting to Ire1 foci. **a**, **d**, **f**, **g**, Localization of Ire1–mCherry as well as either SpR^{U1A} mRNA (SpR WT), as in Fig. 2i, or an intron-less variant (SpR Δintron, **a**), of *PGK1*^{U1A} bearing either the wild-type (*PGK1*-3′*hac1*, **d**, **f**) or the mutant Δ3′ BE (*PGK1*-3′ *hac1*Δ3′ BE, **g**) 3′ UTR stem-loop of *HAC1* mRNA, in combination with (**f**, **g**) or without (**d**) a small stem-loop (SL) that confers translational repression in its 5′ UTR, as schematically depicted. **b**, Co-localization index for mRNA recruitment of WT and Δintron splicing reporter variants into Ire1 foci (means and s.e.m., $n = 5$); the bar for the Δ3′ BE mutant as depicted in Fig. 2i is shown for comparison. **c**, Northern blot of *HAC1* mRNA from yeast strains that overexpressed variants of the splicing reporter, as indicated. **e**, Western blot of the variants of HA-tagged Pgk1 protein (HA–Pgk1) bearing the 3′ UTR from *HAC1* with or without a 5′ UTR stem-loop (SL). **a**, **c**–**g**, ER stress was induced with 10 mM DTT for 45 min. **h**, Co-localization index for mRNA recruitment into Ire1 foci of *PGK1*^{U1A} wild type (see Fig. 2f) or variants shown in **d**, **f** and **g** (means and s.e.m., $n = 5$–8).

therefore, is not only key to facilitate timely synthesis of Hac1 protein on induction of the UPR, but is also integral to the targeting of *HAC1* mRNA to ER stress signalling centres.

Our results describe the first example, to our knowledge, of mRNA targeting as a central feature in a signalling pathway. *HAC1* mRNA is delivered to the site where it is processed as part of the main switch regulating the UPR. The mRNA guidance mechanisms characterized so far serve other goals, such as delivery of mRNA to sites of storage or degradation[18,19], or restricted distribution of the proteins they encode[20–22]. *HAC1* mRNA delivery to Ire1 foci has in common with other mRNA-targeting mechanisms that it depends on a signal in the 3′ UTR and on translational repression of the mRNA[23]. The mechanism of translational control of *HAC1* mRNA serves both to prevent translation of a functional transcription factor when the UPR is off, and to allow the mRNA access to the splicing machine, which removes the intron to allow its translation, when the UPR is on. In this way, the targeting signal is inactivated when translation of *HAC1* mRNA resumes, even though the 3′ BE remains present in the spliced mRNA.

The translational block in *Saccharomyces cerevisiae* is exerted by means of a 16-base-pairing interaction between sequences in the 252-nucleotide-long intron and the 5′ UTR[4]. Most *HAC1* or XBP1 orthologues bear introns that are shorter (~20–26 nucleotides) and show no sequence complementarity to support 5′ UTR/intron-based translational blocks. It is conceivable that other means of translational repression come into play. For instance, the general translational attenuation in response to ER stress as mediated by the ER-resident transmembrane eIF2α kinase PERK[24] could serve a functionally similar role in XBP1 mRNA targeting in metazoans.

Our findings emphasize the role of Ire1 oligomers, rather than dimers, in UPR signalling. Early co-immunoprecipitation studies already provided evidence for oligomerization[25], and the identification of two functionally important interfaces that link Ire1 luminal domains into linear filaments in the crystal lattice supports an attractive model by which neighbouring Ire1 molecules are 'stitched' together by the binding of unfolded proteins in the ER lumen[15]. This model and the epistasis data in Fig. 2h indicate that Ire1 foci formation is governed by self-organization. Overexpression of Ire1 caused an enlargement of the foci, but did not increase their number (not shown), indicating that there is a limited number of nucleation sites per cell and that foci may arise at such predisposed sites at the ER membrane. Because *HAC1* mRNA recruitment occurs with amazing speed and efficiency (for example, Fig. 3), one can further speculate that the 3′ BE-containing targeting signal may allow *HAC1* mRNA to travel actively along cytoskeletal filaments to these pre-disposed sites, where Ire1 concentrates.

Clustering of activated signalling receptors occurs in many systems, such as in the immunological synapse[26] and in bacterial chemotaxis[27], and the resulting local concentration of the signalling machinery can greatly enhance the efficiency of signal transduction. Interestingly, we found that on oligomerization *in vitro* the nuclease activity of the Ire1 kinase/nuclease domains vastly increases[28] over the activity observed for Ire1 dimers[29]. Thus, by clustering into oligomers, Ire1 acquires enhanced avidity towards its substrate *HAC1* mRNA and reaches full enzymatic activation at the same time. These mechanistic features converge into a signalling relay that provides the efficiency and timeliness required to combat ER stress.

METHODS SUMMARY

Microscopy data acquisition and analysis. Cells were visualized on a Yokogawa CSU-22 spinning disc confocal on a Nikon TE2000 microscope. Images of Ire1–mCherry and U1A–GFP-decorated *HAC1*^{U1A}, SpR^{U1A} and *PGK1*^{U1A} mRNAs and variants thereof were analysed using a customized MatLab script to determine the fraction of Ire1–mCherry in foci and to score the recruitment of U1A–GFP-decorated mRNA in Ire1 foci. The annotated MatLab script is available in the Supplementary Information. In brief, after background subtraction we defined the fraction of Ire1–mCherry in foci as the ratio between the integrated fluorescence intensity of pixels with a signal greater than a threshold value and the total integrated fluorescence intensity. The threshold was empirically defined such that under non-stress conditions no signal was scored as 'foci'. Similarly, RNA foci were defined as pixels exceeding by twofold the mean intensity in the RNA channel. A 'co-localization index' was then defined as the integrated intensity of the pixels within the RNA foci that had pixels in common with Ire1 foci divided by the total RNA intensity, and expressed in arbitrary units in a range from 0 to 100. For each condition, the percentage of Ire1–mCherry in foci and the co-localization index for the mRNA recruited to the foci was determined for 5–9 individual cells. Values and the standard error of the mean are given in histograms in Figs 2–4. Because, in contrast to the covalently fluorescently tagged Ire1, we do not know what fraction of U1A–GFP is bound to mRNAs containing U1A-binding sites, background subtraction for U1A–GFP was arbitrary. Therefore, we quantified the data by the co-localization index rather an absolute percentage co-localization measure. The co-localization index robustly scores the differences in mRNA recruitment we observed qualitatively in the fluorescent micrographs.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 20 August; accepted 5 November 2008.
Published online 14 December 2008.

1. Bernales, S., Papa, F. R. & Walter, P. Intracellular signaling by the unfolded protein response. *Annu. Rev. Cell Dev. Biol.* **22,** 487–508 (2006).
2. Ron, D. & Walter, P. Signal integration in the endoplasmic reticulum unfolded protein response. *Nature Rev. Mol. Cell Biol.* **8,** 519–529 (2007).
3. van Anken, E. & Braakman, I. Endoplasmic reticulum stress and the making of a professional secretory cell. *Crit. Rev. Biochem. Mol. Biol.* **40,** 269–283 (2005).
4. Rüegsegger, U., Leber, J. H. & Walter, P. Block of *HAC1* mRNA translation by long-range base pairing is released by cytoplasmic splicing upon induction of the unfolded protein response. *Cell* **107,** 103–114 (2001).
5. Sidrauski, C., Cox, J. S. & Walter, P. tRNA ligase is required for regulated mRNA splicing in the unfolded protein response. *Cell* **87,** 405–413 (1996).
6. Sidrauski, C. & Walter, P. The transmembrane kinase Ire1p is a site-specific endonuclease that initiates mRNA splicing in the unfolded protein response. *Cell* **90,** 1031–1039 (1997).
7. Travers, K. J. *et al.* Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation. *Cell* **101,** 249–258 (2000).
8. Gonzalez, T. N., Sidrauski, C., Dörfler, S. & Walter, P. Mechanism of nonspliceosomal mRNA splicing in the unfolded protein response pathway. *EMBO J.* **18,** 3119–3132 (1999).
9. Brodsky, A. S. & Silver, P. A. Pre-mRNA processing factors are required for nuclear export. *RNA* **6,** 1737–1749 (2000).
10. Diehn, M., Eisen, M. B., Botstein, D. & Brown, P. O. Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. *Nature Genet.* **25,** 58–62 (2000).
11. Kimata, Y. *et al.* Two regulatory steps of ER-stress sensor Ire1 involving its cluster formation and interaction with unfolded proteins. *J. Cell Biol.* **179,** 75–86 (2007).
12. Brengues, M., Teixeira, D. & Parker, R. Movement of eukaryotic mRNAs between polysomes and cytoplasmic processing bodies. *Science* **310,** 486–489 (2005).
13. Teixeira, D. & Parker, R. Analysis of P-body assembly in *Saccharomyces cerevisiae*. *Mol. Biol. Cell* **18,** 2274–2287 (2007).
14. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425,** 737–741 (2003).
15. Credle, J. J., Finer-Moore, J. S., Papa, F. R., Stroud, R. M. & Walter, P. On the mechanism of sensing unfolded protein in the endoplasmic reticulum. *Proc. Natl Acad. Sci. USA* **102,** 18773–18784 (2005).
16. Pollock, R. & Rivera, V. M. Regulation of gene expression with synthetic dimerizers. *Methods Enzymol.* **306,** 263–281 (1999).
17. Chartrand, P., Meng, X. H., Huttelmaier, S., Donato, D. & Singer, R. H. Asymmetric sorting of Ash1p in yeast results from inhibition of translation by localization elements in the mRNA. *Mol. Cell* **10,** 1319–1330 (2002).
18. Anderson, P. & Kedersha, N. RNA granules. *J. Cell Biol.* **172,** 803–808 (2006).
19. Parker, R. & Sheth, U. P bodies and the control of mRNA translation and degradation. *Mol. Cell* **25,** 635–646 (2007).
20. Kindler, S., Wang, H., Richter, D. & Tiedge, H. RNA transport and local control of translation. *Annu. Rev. Cell Dev. Biol.* **21,** 223–245 (2005).
21. Choi, S. B. *et al.* Messenger RNA targeting of rice seed storage proteins to specific ER subdomains. *Nature* **407,** 765–767 (2000).
22. Takizawa, P. A., DeRisi, J. L., Wilhelm, J. E. & Vale, R. D. Plasma membrane compartmentalization in yeast by messenger RNA transport and a septin diffusion barrier. *Science* **290,** 341–344 (2000).
23. Czaplinski, K. & Singer, R. H. Pathways for mRNA localization in the cytoplasm. *Trends Biochem. Sci.* **31,** 687–693 (2006).
24. Harding, H. P., Zhang, Y. & Ron, D. Protein translation and folding are coupled by an endoplasmic-reticulum-resident kinase. *Nature* **397,** 271–274 (1999).
25. Shamu, C. E. & Walter, P. Oligomerization and phosphorylation of the Ire1p kinase during intracellular signaling from the endoplasmic reticulum to the nucleus. *EMBO J.* **15,** 3028–3039 (1996).
26. Bromley, S. K. *et al.* The immunological synapse. *Annu. Rev. Immunol.* **19,** 375–396 (2001).
27. Maddock, J. R. & Shapiro, L. Polar location of the chemoreceptor complex in the *Escherichia coli* cell. *Science* **259,** 1717–1723 (1993).
28. Korennykh, A. V. *et al.* The unfolded protein response signals through high-order assembly of Ire1. *Nature* doi:10.1038/nature07661 (this issue).
29. Lee, K. P. *et al.* Structure of the dual enzyme Ire1 reveals the basis for catalysis and regulation in nonconventional RNA splicing. *Cell* **132,** 89–100 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Jonikas and B. Kornmann for their help with the MatLab scripts; R. Parker for the pPS2037 and pRP1187 plasmids; K. Thorn for the pKT127 plasmid and for his assistance with microscopy at the Nikon Imaging Center at UCSF; and C. Guthrie, R. Andino, J. Gross and members of the Walter laboratory for discussion and comments on the manuscript. T.A. was supported by the Basque Foundation for Science and the Howard Hughes Medical Institute; E.v.A. by the Netherlands Organization for Scientific Research (NWO); D.P. and C.A.R. by the National Science Foundation; C.A.R. by the President's Dissertation Year Fellowship; and A.V.K. by the Jane Childs Memorial Fund for Medical Research. P.W. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions T.A. and E.v.A. wrote the manuscript, conceived the experiments and together with D.P. carried out most of the experimental work. I.M.S. and E.v.A. observed Ire1 foci, and C.A.R. performed all experiments concerning tRNA ligase localization. A.V.K. carried out kinetic analyses. P.W. directed the research programme and writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.A. (Tomas.Aragon@ucsf.edu).

METHODS

Yeast strains and plasmids. Standard cloning and yeast techniques were used for construction, transformation and integration of plasmids[30–32]. HA-tagged versions of *HAC1* with either its own 3′ UTR or that of *ACT1* or *PGK1* were integrated as a genomic copy, replacing endogeneous *HAC1*. The splicing reporter construct was generated by replacing positions 1 to 648 of the *HAC1* coding sequence in exon 1 with the GFP ORF. In the Δ3′ BE mutants, positions 176–182 and 212–218 of the 3′ UTR of *HAC1* were deleted. The 3′ BE stem that was placed between the stop codon and the *ACT1* 3′ UTR of the splicing reporter comprised positions 155–187 and 207–236 of the 3′ UTR of *HAC1*. The mRNA visualization constructs were created by inserting into the pRS426 vector[33] the sequences of *PGK1*, *HAC1* or a non-fluorescent GFP–R96A mutant of the splicing reporter ending at position 280 of the 3′ UTR of *HAC1*, followed by 16 tandem repeats of the U1A binding sequence and the *PGK1* terminator, derived from pPS2037 (a gift from R. Parker), and a polyA signal. A copy of the U1A RNA-binding domain fused to GFP was integrated into the genome from plasmid pRP1187 (a gift from R. Parker). Surprisingly, the key to the low noise in the imaging lies in the curious fact that in pRP1187 the U1A–GFP ORF is inserted backwards, so that its expression is driven by a cryptic, uncharacterized promoter element within the (reverse) *PGK1* transcription terminator. The low levels of U1A–GFP expression derived from this construct prove ideal for mRNA imaging. By PCR, a previously described[17] 5′ stem-loop structure was introduced 26-nucleotides upstream of the start codon of *PGK1*, and nucleotides 108–280 of the *HAC1* 3′ UTR, comprising the entire stem, were inserted after the *PGK1* stop codon, where indicated. A monomeric (A206R), yeast-codon-adapted version of GFP, derived from pKT127[34], or mCherry was placed into Ire1 between residues I571 and G572, and the FKBP-derived Fv2E domain (Ariad) between R112 and Y449, replacing the core ER-stress-sensing domain[15]. Ire1 luminal interface mutants are: $IF1^{L}$ (T226W/F247A), $IF2^{L}$ (W426A) and $IF1/2^{L}$ (T226W/F247A/W426A)[15]. Ire1 variants in all assays were expressed at near-endogenous levels from centromeric pRS315.

RNA and protein analysis. RNA preparation, electrophoresis, labelling of probes for northern blot analysis and quantification of splicing efficiencies were performed as described[4]. Protein extraction, electrophoresis and transfer to nitrocellulose for immunoblot analysis with anti-HA antibody were performed as described[4].

Microscopy. All samples were taken from yeast cells that were kept in early log phase for at least 24 h in synthetic media containing excess amounts of adenine and tryptophan before imaging. Light microscopy was done with a Yokogawa CSU-22 spinning disc confocal on a Nikon TE2000 microscope. GFP was excited with the 488 nm Ar-ion laser line and mCherry with the 568 nm Ar-Kr laser line. Images were recorded with a ×100/1.4 NA Plan Apo objective on a Cascade II EMCCD. The sample magnification at the camera was 60 nm per pixel. The microscope was controlled with μManager and ImageJ. Images were selected for analysis and for display in figures to contain no saturated pixels (in case of the RNA imaging) and a signal substantially above background (in case of Ire1–mCherry imaging). We excluded images of cells with strong vacuolar autofluorescence. Images were processed in ImageJ and Adobe Photoshop such that the linear range of the signal was comparable between images.

Quantitative analysis of Ire1 foci and co-localization of mRNA in foci. Images of Ire1–mCherry and U1A–GFP-decorated $HAC1^{U1A}$, SpR^{U1A} and $PGK1^{U1A}$ mRNAs and variants thereof were analysed using a customized MatLab script to determine the fraction of Ire1–mCherry in foci and to score the recruitment of U1A–GFP-decorated mRNA in Ire1 foci. The annotated MatLab script is available (Supplementary Information). In brief, the mean pixel intensity of a background area (~20% of section area) was defined in an intracellular area excluding ER. Ire1 was defined as all signal exceeding the mean background by 1.1-fold. Under non-stress conditions, we never observed this signal to exceed a 1.5-fold background threshold. We thus defined the fraction of Ire1 in foci as the ratio of Ire1–Cherry fluorescence intensity above threshold divided by the total Ire1–Cherry fluorescence intensity. The threshold was empirically defined such that under non-stress conditions no signal was scored as 'foci'. Similarly, RNA foci were defined as pixels exceeding by twofold the mean intensity in the RNA channel. A 'co-localization index' was then defined as the integrated intensity of the pixels within the RNA foci that had pixels in common with Ire1 foci divided by the total RNA intensity, and expressed in arbitrary units in a range of 0 to 100. For each condition, the percentage of Ire1–mCherry in foci and the co-localization index for the mRNA recruited to the foci was determined for 5–9 individual cells. Values and the standard error of the mean are given in histograms in Figs 2–4. Because, in contrast to the covalently fluorescently tagged Ire1, we do not know what fraction of the fluorescent reporter U1A–GFP in cells is bound to mRNAs containing U1A binding sites, background subtraction for U1A–GFP was arbitrary. Therefore, we report co-localization by this 'co-localization index' rather than by an absolute percentage co-localization measure. The co-localization index robustly scores the differences in mRNA recruitment we observed in the fluorescent micrographs.

30. Guthrie, C. & Fink, G. R. *Guide to Yeast Genetics and Molecular and Cell Biology* (Academic, 2002).
31. Longtine, M. S. *et al.* Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae. Yeast* **14**, 953–961 (1998).
32. Sambrook, J., Maniatis, T. & Fritsch, E. F. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, 1989).
33. Sikorski, R. S. & Hieter, P. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae. Genetics* **122**, 19–27 (1989).
34. Sheff, M. A. & Thorn, K. S. Optimized cassettes for fluorescent protein tagging in *Saccharomyces cerevisiae. Yeast* **21**, 661–670 (2004).

LETTERS

Peptide neurotransmitters activate a cation channel complex of NALCN and UNC-80

Boxun Lu[1]*, Yanhua Su[1]*†, Sudipto Das[1], Haikun Wang[1], Yan Wang[1], Jin Liu[1] & Dejian Ren[1]

Several neurotransmitters act through G-protein-coupled receptors to evoke a 'slow' excitation of neurons[1,2]. These include peptides, such as substance P and neurotensin, as well as acetylcholine and noradrenaline. Unlike the fast (approximately millisecond) ionotropic actions of small-molecule neurotransmitters, the slow excitation is not well understood at the molecular level, but can be mainly attributed to suppressing K^+ currents and/or activating a non-selective cation channel[3–9]. The molecular identity of this cation channel has yet to be determined; similarly, how the channel is activated and its relative contribution to neuronal excitability induced by the neuropeptides are unknown. Here we show that, in the mouse hippocampal and ventral tegmental area neurons, substance P and neurotensin activate a channel complex containing NALCN and a large previously unknown protein UNC-80. The activation by substance P through TACR1 (a G-protein-coupled receptor for substance P) occurs by means of a unique mechanism: it does not require G-protein activation but is dependent on Src family kinases. These findings identify NALCN as the cation channel activated by substance P receptor, and suggest that UNC-80 and Src family kinases, rather than a G protein, are involved in the coupling from receptor to channel.

NALCN is a neuronal cation channel carrying a small background leak Na^+ current at the resting membrane potential[10]. When overexpressed in HEK293T fibroblast cells, it generates a sodium ion (Na^+)-permeable cation channel that is voltage-independent, non-inactivating, tetrodotoxin (TTX)-resistant and gadolinium ion (Gd^{3+})-blockable[10]. It is not known whether, like background potassium ion (K^+) channels, NALCN is also regulated by neuromodulators, but the biophysical and pharmacological properties of the NALCN currents (I_{NALCN}) closely resemble those of the substance P (SP)-activated cation channel currents (I_{SP}) studied in several brain regions[11–14].

To test the possibility that I_{SP} requires NALCN, we recorded I_{SP} in wild-type and *Nalcn* knockout[10] (*Nalcn*$^{-/-}$) mouse neurons by means of patch-clamp recording with measures taken to minimize K^+ channel effects and to block voltage-gated Na^+ channel and synaptic currents. In 16 out of 34 wild-type hippocampal pyramidal neurons held at −67 mV, an inward current (>50 pA) developed within 1.1 ± 0.2 min (range of 0.4–2.9 min) after a pulse of SP was delivered through a puffer pipette placed ~20 μm from the cell body (Fig. 1a, e). The time courses of I_{SP} vary between neurons, but are comparable to previous recordings[11,12,15] and are in agreement with the 'slow' nature of the signal transduction pathway (see below). I_{SP} was blocked by pre-incubating cells with a competitive peptide ($n = 3$; Fig. 1b) or a TACR1 antagonist (L703606; $n = 14$; Fig. 1a), suggesting involvement of TACR1 or another member of the tachykinin receptor family of G-protein-coupled receptors (GPCRs). Reducing the bath concentration of Na^+ from 155 mM to 5 mM largely abolished the I_{SP} (Fig. 1a), suggesting that Na^+ was the main charge carrier of the current at this holding potential. I_{SP} has also been recorded in the rat ventral tegmental area (VTA)[9]. Out of 43 putative dopaminergic neurons cultured from the VTA of wild-type mice (see Methods), 32 had an $I_{SP} > 50$ pA (Fig. 1d). In contrast to the wild-type mice, none of the 30 hippocampal (Fig. 1c, e) and 29 VTA (Fig. 1d) neurons from *Nalcn*$^{-/-}$ mutant mice showed a significant I_{SP} (Fig. 1e).

Like I_{NALCN}[10], the SP-activated currents did not inactivate within 300 ms at any voltage (Fig. 2a). The averaged current–voltage (*I–V*) relationship was linear at negative membrane potentials, suggesting

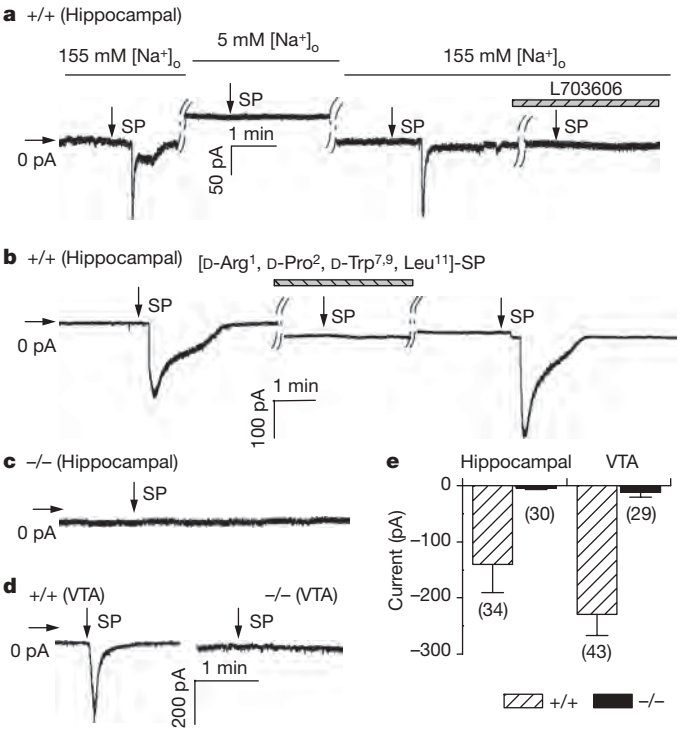


Figure 1 | NALCN is required for I_{SP}. Currents (at −67 mV) were recorded from wild-type (+/+) or *Nalcn* knockout (−/−) hippocampal and VTA neurons. Horizontal and vertical arrows indicate 0 current level and SP application (10 s of puffing), respectively. **a**, I_{SP} developed in bath containing 155 mM Na^+, but not when Na^+ was lowered to 5 mM (replaced with NMDG). Incubation with L703606 (10 μM, 6 min) blocked I_{SP}. **b**, Blockade by a peptide TACR1 antagonist ([D-Arg^1, D-Pro^2, D-Trp^{7,9}, Leu^{11}]-SP; 10 μM, 6 min incubation). **c**, *Nalcn*$^{-/-}$ hippocampal neuron. **d**, VTA neurons. **e**, Summary of the I_{SP} sizes. Numbers of cells are in parentheses. Error bars, mean and s.e.m.

[1]Department of Biology, University of Pennsylvania, 415 S. University Avenue, Philadelphia, Pennsylvania 19104, USA. †Present address: State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China.
*These authors contributed equally to this work.

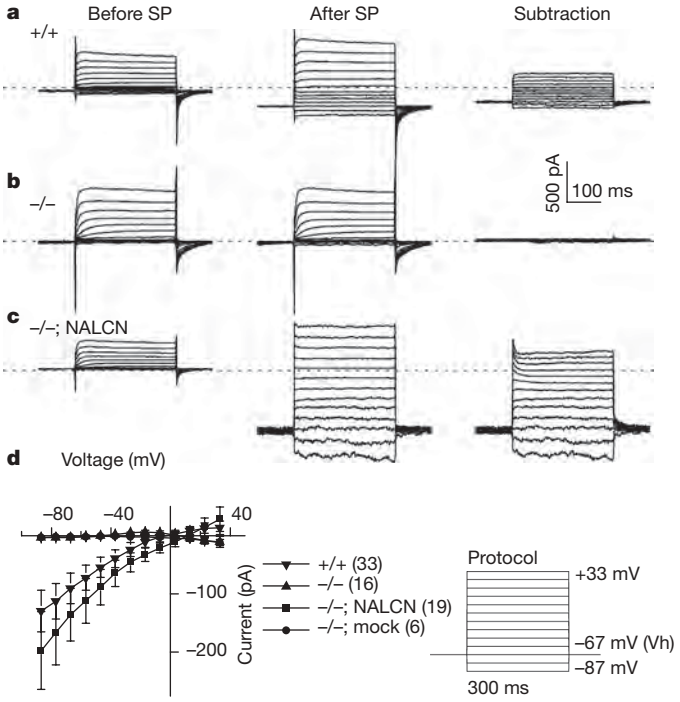


Figure 2 | Characterization of the I_{SP} in hippocampal neurons. **a**–**c**, Net SP-activated currents at various voltages (right) were obtained by subtracting the currents before (left) from those after (middle) SP bath application (1 μM) in wild type (+/+, **a**), mutant (−/−, **b**) and mutant transfected with NALCN (−/−; NALCN, **c**). Dotted lines indicate 0 current level. **d**, Averaged I_{SP} amplitudes. The lines from mutant (−/−) and mutant transfected with empty vector (−/−; mock) overlap. Numbers of cells are in parentheses. The right panel shows the voltage step protocols used. Vh, holding voltage. Error bars, mean ± s.e.m.

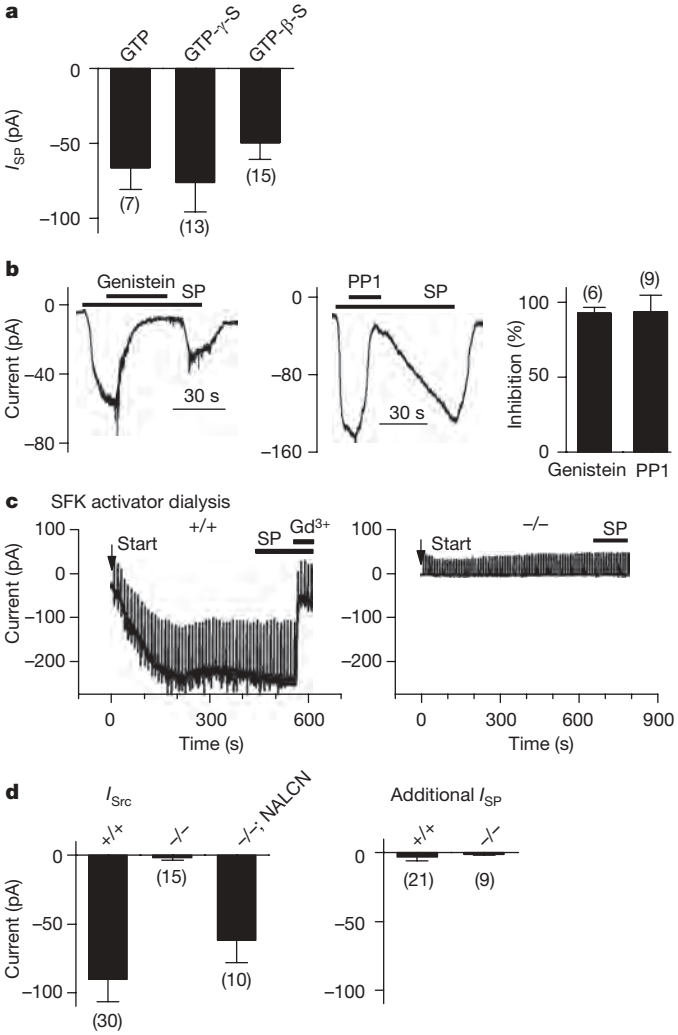


Figure 3 | I_{SP} is G-protein-independent but requires SFKs. Hippocampal neurons were used. **a**, I_{SP} (at −67 mV) from recordings with GTP-, GTP-γ-S- and GDP-β-S-containing pipette solutions. Numbers of cells are in parentheses. **b**, Inhibition of I_{SP} (at −67 mV) by genistein (30 μM; left) and PP1 (20 μM; middle). Right, summary. **c**, In wild-type neurons (left), a Gd^{3+}-blockable current developed after intracellular dialysis (start time indicated by arrow) with pipette solution containing an SFK activator (1 μM; a ramp from −67 to −47 mV in 1.4 s was given every 10 s to monitor input resistance). After the current reached a plateau (size defined as I_{Src} for **d**), SP did not induce an additional current (I_{SP} for **d**). Right, *Nalcn*$^{-/-}$ neuron. **d**, Summary of I_{Src} (left) and additional currents activated by SP after the dialysis-induced currents plateaued (right). Error bars, mean and s.e.m.

little voltage-dependence of conductance in this range (Fig. 2d). The overall properties of the I_{SP} in cultured mouse hippocampal neurons were similar to those of the I_{SP} recorded from numerous other neuronal preparations, primarily from brain slices[12,13,16]. Owing to low current amplitudes, potential space-clamping problems in neuronal recordings, and possible contamination by voltage-gated currents at positive potentials, a detailed biophysical characterization could not be performed in the neurons.

I_{SP} could be restored in the *Nalcn*$^{-/-}$ neurons by transfecting a *Nalcn* complementary DNA (Fig. 2c, d). The restored I_{SP} (Supplementary Fig. 2b, d), like that from wild-type neurons (Supplementary Fig. 2a, d), was blocked by a trivalent NALCN blocker Gd^{3+} (10 μM). The current, however, became resistant to 10 μM Gd^{3+} (Supplementary Fig. 2c, d) when restored with a Gd^{3+}-resistant NALCN pore mutant (EEKA) that has a single amino acid change (E to A) in the ion filter region (Supplementary Fig. 1). Taken together, the similarity of I_{SP} and I_{NALCN}, the absence of I_{SP} in *Nalcn*$^{-/-}$ neurons and presence with *Nalcn* cDNA transfection, and the alteration of I_{SP} pharmacology by the EEKA pore mutant strongly suggest that NALCN forms the pore conduit carrying the I_{SP}.

To determine the role of G proteins (the immediate downstream effectors of the GPCR TACR1) in the NALCN activation by SP, we 'locked' G proteins in active or inactive states with non-hydrolysable analogues of GTP (GTP-γ-S) or GDP (GDP-β-S), respectively, applied by means of patch pipettes. Surprisingly, SP still induced inward currents of comparable sizes (Fig. 3a). Thus, the activation of NALCN by SP through GPCRs is probably by means of an unconventional mechanism that does not require G-protein stimulation.

Some GPCRs may also activate the Src family of tyrosine kinases (SFKs)—a more recently discovered GPCR signalling cascade that regulates downstream factors such as mitogen-activated protein kinase and gene expression[17,18]. Bath application of genistein (a phosphotyrosine kinase inhibitor) or PP1 (an SFK inhibitor) abolished I_{SP} (Fig. 3b), suggesting that SFKs are required for the activation of NALCN by SP.

Similar to previous findings[19], intracellular dialysis with an SFK activator by means of a patch pipette induced a gradual increase of inward current (defined as I_{Src}; Fig. 3c, left panel). After the current plateaued, SP no longer activated an additional inward current (I_{SP}; Fig. 3c, d), suggesting that SP and SFKs activate a common channel. Similar to I_{SP}, the I_{Src} was blocked by Gd^{3+} (Fig. 3c, left). In contrast to wild-type neurons, *Nalcn*$^{-/-}$ neurons lacked I_{Src} (Fig. 3c, d); this current was largely restored by transfection with *Nalcn* cDNA (Fig. 3d). These data suggest that SFK activation is both necessary and sufficient for I_{SP}. Given that many other ion channels can also be regulated by SFKs[20], the lack of I_{Src} in the *Nalcn*$^{-/-}$ neurons was surprising, but seems to suggest that NALCN is a major cation channel target for SFKs near the resting membrane potential and may also be modulated by the diverse array of stimuli (for example, neurotransmitters, growth factors, cytokines, cell stretch and adhesion molecules) that lead to SFK activation[21,22].

Other neuropeptides such as neurotensin (NT) also elicit slow depolarization of neurons and activate a cation current (I_{NT}) similar to I_{SP}[9]. In the wild-type VTA neurons, NT puffing (10 μM) elicited an inward current (-90.0 ± 25.3 pA at -67 mV; $n = 29$) that was blocked by an NT receptor antagonist SR48692 and the SFK inhibitor PP1 (not shown). In contrast, neurons from *Nalcn*$^{-/-}$ mice lacked an I_{NT} (Supplementary Fig. 3), suggesting that the I_{NT} is also through NALCN.

To determine the role of the SP- and NT-activated NALCN currents in modulating neuronal excitability, we recorded action potentials in the VTA dopaminergic neurons. Firing frequencies were significantly increased by 1 μM SP (Supplementary Fig. 4a, c) or 1 μM NT (Supplementary Fig. 4b, c) in the wild-type, but not in the *Nalcn*$^{-/-}$, neurons. The defect in the mutant was partially restored by transfecting with *Nalcn* cDNA (Supplementary Fig. 4c). We conclude that the NALCN channel has a major role in the potentiation of neuronal excitability by the neuropeptides in these neurons. Under other conditions or in other cells, the peptides may also excite neurons by suppressing K^+ currents or by activating some of the TRP family members[6,8,9,23–26].

Unlike in neurons, little robust I_{SP} was observed from HEK293T fibroblast cells co-transfected with TACR1 and NALCN (see Fig. 4e), possibly because of the lack of a key component for the channel activation. In *Drosophila melanogaster* and *Caenorhabditis elegans*, *Nalcn* genetically interacts with other genes such as *Unc-79* and *Unc-80* (refs 27–29). To investigate whether UNC-80 forms a physical complex with NALCN in the brain, we cloned a mammalian UNC-80 homologue (hereafter called mUNC-80) from mouse brain (Supplementary Fig. 5). The full-length mUNC-80 is predicted to encode a 371 kDa protein (Fig. 4a) with no obvious domains of known function. It has high (96%) and moderate (~30%) identities with its human and invertebrate homologues, respectively. NALCN and mUNC-80 form a complex in mouse brain (Fig. 4c), and in transfected HEK293T cells (Fig. 4b). In addition, both are tyrosine-phosphorylated and such phosphorylation can be inhibited by PP1 (Supplementary Fig. 6).

In HEK293T cells with moderate levels of co-expression of TACR1, mUNC-80 and NALCN (see Methods), SP activated a current (-550.6 ± 102.9 pA at -100 mV; 0 to $-4{,}904$ pA; Fig. 4d, e) with a linear *I–V* relationship (Fig. 4d, right, curve 2). Like I_{SP} in neurons, the current was blocked by the NALCN blocker Gd^{3+} (Fig. 4f), but became resistant to Gd^{3+} when NALCN was replaced with the EEKA pore mutant (Fig. 4g). Inclusion of GDP-β-S in the pipette did not prevent current activation (-914.0 ± 317.1 pA; $n = 10$), suggesting independence of G-protein activation.

Consistent with a requirement of SFKs, the current was suppressed by SFK inhibitors (Fig. 4g, h and Supplementary Fig. 7) and an anti-phospho-SFK antibody (Supplementary Fig. 8). Furthermore, when a constitutively active Src (with Y529 mutated to F) was co-transfected, a basal current was increased from -104.1 ± 22.1 pA (empty vector to replace Src Y529F in transfection; $n = 22$) to -434.3 ± 119.9 pA ($n = 21$), and application of SP no longer activated significant I_{SP} in the Src Y529F co-transfected cells (-64.6 ± 33.1 pA, $n = 13$, compared with -416.2 ± 89.6 pA in the empty vector co-transfection control, $n = 13$; Supplementary Fig. 9). Finally, inclusion of a recombinant active Src protein in the pipette led to a gradual increase of inward current (Supplementary Fig. 10), suggesting that, like in neurons (Fig. 3c), SFK is sufficient to activate the current.

The simplest interpretation of our work is that the SP- and NT-activated cation current is through a channel complex consisting of NALCN and mUNC-80. Our preliminary studies found that UNC-79 is also associated with mUNC-80 and NALCN in the brain (not shown). Further studies need to investigate whether UNC-79 affects the properties of I_{SP} such as the activation time courses. Similarly, the mechanisms underlying the involvement of SFKs in the coupling between receptor and channel remain to be uncovered. Owing to the large driving force of Na^+ at resting membrane potential, I_{NALCN} should be a potent excitatory current in neurons. The channel is voltage-independent and non-inactivating, and is thus ideal for regulating neuronal excitability by neuromodulators.

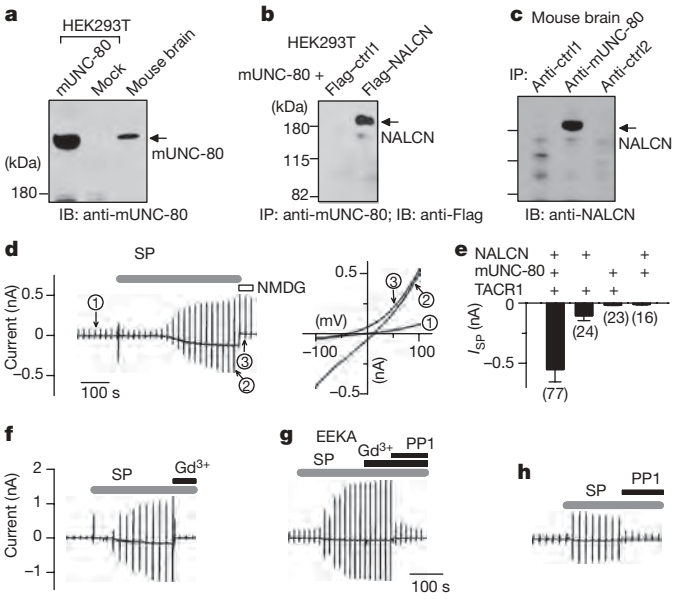


Figure 4 | I_{SP} reconstituted in HEK293T cells. **a**, Immunoblot (IB) with lysates from transfected HEK293T cells and brain. **b**, **c**, Immunoprecipitation (IP) showing the protein complex between mUNC-80 and NALCN in HEK293T cells transfected as indicated (**b**) and in brain (**c**). Ctrl1 and ctrl2 are two unrelated proteins used as controls. Recordings in **d**–**h** were performed using ramp protocols (holding voltage Vh = −20 mV; −100 to +100 mV in 1 s, every 20 s). **d**, Recordings from a cell transfected with TACR1, mUNC-80 and NALCN. Currents at three time points are expanded in the right panel ((1) before SP; (2) after SP; and (3) after Na^+ and K^+ in the bath were replaced with NMDG). **e**, Summary of I_{SP} sizes (at −100 mV) from cells transfected with combinations as indicated. **f**, **g**, Recordings showing that I_{SP} was blocked by Gd^{3+} (**f**) but became resistant to Gd^{3+} when NALCN was replaced by the EEKA pore mutant (**g**). **h**, Inhibition of I_{SP} by PP1 (20 μM). Error bars, mean and s.e.m.

METHODS SUMMARY

For recording the basal current in NALCN-overexpressing HEK293T cells[10] (Supplementary Fig. 1), only cells expressing the highest amount of NALCN (top ~5%; judged by the levels of green fluorescent protein (GFP) expression from the pTracer vector containing both GFP and NALCN under separate promoters) were used. Receptor-activated currents (Fig. 4 and Supplementary Figs 7–10) were recorded from cells with modest level of expression (top ~40%).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 22 June; accepted 23 October 2008.
Published online 17 December 2008.

1. Kandel, E. R., Schwartz, J. H. & Jessell, T. M. *Principles of Neural Science* 229–252 (McGraw-Hill, 2000).
2. Hille, B. *Ion Channels of Excitable Membranes* 201–236 (Sinauer, 2001).
3. Kuba, K. & Koketsu, K. Synaptic events in sympathetic ganglia. *Prog. Neurobiol.* **11**, 77–169 (1978).
4. Jan, Y., Jan, L. & Kuffler, S. Further evidence for peptidergic transmission in smypathetic ganglia. *Proc. Natl Acad. Sci. USA* **77**, 5008–5012 (1980).
5. Kuffler, S. & Sejnowski, T. Muscarinic and peptidergic excitation of bull-frog sympathetic neurons. *J. Physiol. (Lond.)* **341**, 257–278 (1983).
6. Stanfield, P. R., Nakajima, Y. & Yamaguchi, K. Substance P raises neuronal membrane excitability by reducing inward rectification. *Nature* **315**, 498–501 (1985).
7. Shen, K. Z. & North, R. A. Muscarine increases cation conductance and decreases potassium conductance in rat locus coeruleus neurones. *J. Physiol. (Lond.)* **455**, 471–485 (1992).
8. Shen, K. Z. & Surprenant, A. Common ionic mechanisms of excitation by substance P and other transmitters in guinea-pig submucosal neurones. *J. Physiol. (Lond.)* **462**, 483–501 (1993).

9. Farkas, R. H., Chien, P. Y., Nakajima, S. & Nakajima, Y. Properties of a slow nonselective cation conductance modulated by neurotensin and other neurotransmitters in midbrain dopaminergic neurons. *J. Neurophysiol.* **76,** 1968–1981 (1996).
10. Lu, B. *et al.* The neuronal NALCN channel contributes resting sodium permeability and is required for normal respiratory rhythm. *Cell* **129,** 371–383 (2007).
11. Shen, K. Z. & North, R. A. Substance P opens cation channels and closes potassium channels in rat locus coeruleus neurons. *Neuroscience* **50,** 345–353 (1992).
12. Aosaki, T. & Kawaguchi, Y. Actions of substance P on rat neostriatal neurons *in vitro*. *J. Neurosci.* **16,** 5141–5153 (1996).
13. Inoue, K., Nakazawa, K., Inoue, K. & Fujimori, K. Nonselective cation channels coupled with tachykinin receptors in rat sensory neurons. *J. Neurophysiol.* **73,** 736–742 (1995).
14. Pena, F. & Ramirez, J. M. Substance P-mediated modulation of pacemaker properties in the mammalian respiratory network. *J. Neurosci.* **24,** 7549–7556 (2004).
15. Jones, S. W. Muscarinic and peptidergic excitation of bull-frog sympathetic neurones. *J. Physiol. (Lond.)* **366,** 63–87 (1985).
16. Otsuka, M. & Yoshioka, K. Neurotransmitter functions of mammalian tachykinins. *Physiol. Rev.* **73,** 229–308 (1993).
17. Lefkowitz, R. J. & Shenoy, S. K. Transduction of receptor signals by β-arrestins. *Science* **308,** 512–517 (2005).
18. DeFea, K. A. *et al.* The proliferative and antiapoptotic effects of substance P are facilitated by formation of β-arrestin-dependent scaffolding complex. *Proc. Natl Acad. Sci. USA* **97,** 11086–11091 (2000).
19. Heuss, C., Scanziani, M., Gahwiler, B. H. & Gerber, U. G-protein-indenpendent signaling mediated by metabotropic glutamate receptors. *Nature Neurosci.* **2,** 1070–1077 (1999).
20. Davis, M. J. *et al.* Regulation of ion channels by protein tyrosine phosphorylation. *Am. J. Physiol. Heart Circ. Physiol.* **281,** H1835–H1862 (2001).
21. Salters, M. W. & Kalia, L. V. Src kinases: a hub for NMDA receptor regulation. *Nature Rev. Neurosci.* **5,** 317–328 (2004).
22. Heuss, C. & Gerber, U. G-protein-independent signaling by G-protein-coupled receptors. *Trends Neurosci.* **23,** 469–475 (2000).
23. Montell, C., Birnbaumer, L. & Flockerzi, V. The TRP channels, a remarkably functional family. *Cell* **108,** 595–598 (2002).
24. Clapham, D. E. TRP channels as cellular sensors. *Nature* **426,** 517–524 (2003).
25. Oh, E. J., Gover, T. D., Cordoba-Rodriguez, R. & Weinreich, D. Substance P evokes cation currents through TRP channels in HEK293 cells. *J. Neurophysiol.* **90,** 2069–2073 (2003).
26. Bley, K. R. & Tsien, R. W. Inhibition of Ca^{2+} and K^{+} channels in sympathetic neurons by neuropeptides and other ganglionic transmitters. *Neuron* **4,** 379–391 (1990).
27. Jospin, M. *et al.* UNC-80 and the NCA ion channels contribute to endocytosis defects in synaptojanin mutants. *Curr. Biol.* **17,** 1595–1600 (2007).
28. Yeh, E. *et al.* A putative cation channel, NCA-1, and a novel protein, UNC-80, transmit neuronal activity in *C. elegans*. *PLoS Biol.* **6,** e55 (2008).
29. Humphrey, J. A. *et al.* A putative cation channel and its novel regulator: cross-species conservation of effects on general anesthesia. *Curr. Biol.* **17,** 624–629 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Clapham, C. Deutsch, I. Medina, B. Novarro, M. Schmidt and H. Xu for critically reading earlier versions of the manuscript, J. Xia for help with experiments, H. Yu and L. Yue for cDNA constructs, and Sanofi-Aventis for the gift of SR48692. This work was supported, in part, by funding from American Heart Association, the NIH and the University of Pennsylvania Research Foundation.

Author Contributions B.L. did recordings from neurons (Figs 1–3 and Supplementary Fig. 2) and all the HEK293T cells (Fig. 4 and Supplementary Figs 1 and 7–10). Y.S. contributed to neuronal recordings (Figs 1 and 2 and Supplementary Figs 3 and 4). S.D. contributed to work in Fig. 2. H.W., Y.W. and J.L. did the protein work (Fig. 4 and Supplementary Fig. 6). D.R. started the project, designed experiments and developed the cDNA constructs. B.L. and D.R. wrote the paper.

Author Information The sequence of mUNC-80 is deposited in GenBank under accession number FJ210934. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to D.R. (dren@sas.upenn.edu).

METHODS

Cloning of mUNC-80 and antibody generation. The mouse mUNC-80 was cloned from single-stranded brain cDNA as four fragments using PCR with reverse transcription (RT–PCR) with primers designed from partial sequences in NCBI databases. Multiple clones were sequenced from each fragment and clones without mutation were used to assemble the full-length sequence in vector pcDNA3.1(+). The start of the ORF was unambiguously identified by the presence of an in-frame stop codon in the 5′ UTR. The anti-NALCN antibody was generated in rabbit with a glutathione *S*-transferase (GST)-fusion protein containing the last 80 amino acids of NALCN. The mUNC80 polyclonal antibody was generated against the carboxy terminus (Supplementary Fig. 5). Both antibodies were affinity-purified. Immunoprecipitation and western blotting followed previously described methods[30].

Neuronal culture and transfection. Hippocampal neurons were cultured from P0 pups on glia pre-plated 35-mm dishes and coverslips as described previously[10] and were used between 7 days *in vitro* (DIV) and DIV18. The protocol for VTA neuron culture was modified from that established in rat[31]. In brief, the VTA was dissected from P0 pups and digested with papain (12 units per ml) for 30 min with occasional mixing; digestion was stopped with 10% serum. Tissue was then triturated and plated in culture medium (Neurobasal-A) supplemented with 2% B-27, 0.5× Pen/Strep and 1 mM Glutamax. Owing to the small size of P0 mice, the VTA neuron culture also probably contained neurons from adjacent brain areas. Putative dopaminergic neurons were confirmed by tyrosine hydroxylase staining and were morphologically identified[31–33]. VTA neurons of ages DIV18 to DIV30 were used for patch-clamp analysis. For transfection with Lipofectamine 2000 (Invitrogen), younger neurons (hippocampal, DIV5–DIV7; VTA, DIV6–DIV7) were used because of their higher efficiency. Transfected neurons were used 48–60 h later.

Patch-clamp analysis. All experiments were performed at room temperature (20–25 °C). For recording the basal leak current in HEK293T cells transfected with NALCN alone[10] (Supplementary Fig. 1), cells were transfected with 3 μg NALCN (in a pTracer vector expressing GFP under a separate promoter). Only the most fluorescent cells (~5% of all green ones) that presumably expressed the highest level of NALCN were selected. For recording SP-activated currents, NALCN (0.5 μg) was co-transfected with mUNC-80 (0.5 μg, in pcDNA3.1(+) vector) and human TACR1 (2 μg, in pcDNA3 vector), unless otherwise stated. An empty vector was added to ensure that the same amount of DNA was transfected when one or more constructs were not included. Cells with an above average level of fluorescence (~40% of the total number of green cells) were selected for analysis. It is possible that some of the basal 'leak' currents in neurons and overexpressing HEK293T cells were a result of basal level of receptor activation and tyrosine kinase activity in the cells. Cells with non-specific leak (for example, due to cell damage) were identified by replacing Na^+ and K^+ with NMDG or by application of blockers. In TACR1–mUNC-80–NALCN transfected cells, ~70% (53 out of 77) of those analysed had SP-activated currents >100 pA (at −100 mV). The absence of currents in the rest presumably reflects the efficiency of having all three proteins (two of which are very large) expressed in the same cell, or the varying levels of endogenous signalling molecules. For current amplitude averages, all cells (including the ones without current) were included. HEK293T cells cultured with several batches of sera did not yield robust currents unless they were serum-starved (with Opti-MEM medium) for 9–16 h before recording (not shown). Data from these batches of transfections were not included for analysis.

Standard pipette solutions used for HEK293T cells contained (in mM): 150 Cs, 120 methanesulphonate, 10 NaCl, 10 EGTA, 2 Mg-ATP, 4 $CaCl_2$, 0.3 Na_2GTP and 10 HEPES (pH 7.4, osmolality ~300 mosM). Bath was a Tyrode's solution containing (in mM): 150 NaCl, 3.5 KCl, 1 $MgCl_2$, 1.2 $CaCl_2$, 10 HEPES and 20 glucose (pH 7.4 with NaOH; final Na^+ 155 mM; ~320 mosM). Some cells were patched with pipette solutions containing no GTP; no differences were observed, consistent with the independence of I_{SP} from G-protein activation. GDP-β-S-containing pipette solutions contained 1 mM GDP-β-S and no GTP. As a control, intracellular dialysis with GDP-β-S-containing pipette solutions blocked the activation of TRPC6 current by carbochol, which is G-protein-dependent, in HEK293T cells co-transfected with TRPC6 and m3AChR3 (not shown). When pipette solutions containing anti-phospho-SFK (from Millipore) or recombinant active Src protein (from Stressgen) were used, heat-inactivated (100 °C for 30 min) proteins were used as a control. Storage buffer of the recombinant protein was exchanged for pipette solution with a dilution of ~30,000 times by spinning three times in a concentrator (Microcon-50, Millipore).

Pyramidal hippocampal neurons and presumably dopaminergic VTA neurons used in patch-clamp recordings were morphologically identified[31–33]. Unless otherwise stated, pipette solutions used for neuronal I_{SP} and I_{NT} recordings contained (in mM): 120 CsCl, 4 EGTA, 2 $CaCl_2$, 2 $MgCl_2$, 10 HEPES, 4 Mg-ATP, 0.3 Tris-GTP and 14 phosphocreatine (di-tris salt) (pH adjusted to 7.4 with CsOH; final $[Cs^+]$ 143 mM; free $[Ca^{2+}]$ ~60 nM; 300 mosM). When GTP-γ-S (1.5 mM) or GDP-β-S (1 mM) was included in pipette solution, GTP was omitted and cells were dialysed for 6–9 min before stimulus application. In experiments with SFK-activator-containing pipette solution, 1 μM SFK activator (a tyrosine-phosphorylated peptide that binds to the SH2 domain of Src kinases, from Santa Cruz Biotechnology Inc., catalogue number sc-3052) was added. TTX (0.8 or 1 μM) was added in the bath of Tyrode's solution. For whole bath SP application, concentrated SP (5 mM) was diluted to 50 μM with bath solution and pipetted into the bath to generate a final concentration of ~1 μM. Currents were continuously recorded for 10–20 min on SP application and the peak currents were used to plot the *I–V* curves. In puffer applications, diluted SP (10 μM) was pressure-applied using a pneumatic picopump for 10 s with a glass pipette (~3–5 μm opening) placed ~20 μm away from the neuron.

For current clamp with the VTA neurons (Supplementary Fig. 4), Tyrode's solution was used as the bath; the pipette solution contained (in mM): 135 K-Asp, 5 NaCl, 5 KCl, 1 $MgCl_2$, 1 EGTA, 10 HEPES, 4 Mg-ATP, 0.3 Tris-GTP and 14 phosphocreatine (di-tris) (pH adjusted to 7.4 with KOH; total K^+, 147 mM). Neurons were isolated during current clamp with DL-2-amino-5-phosphonovaleric acid (APV, 10 μM), bicuculline (20 μM) and 6-cyano-7-nitroquinoxaline-2,3-dione disodium salt (CNQX, 20 μM). Some cultured neurons from both the wild type and mutant showed spontaneous firing at 0 holding current. For others, small currents (wild-type, -2.2 ± 4.6 pA, $n = 29$; mutant, $+10.2 \pm 4.4$ pA, $n = 19$) were injected to artificially elicit repetitive firing. Firing frequencies were calculated from time windows (5 min) before and 30 s after SP or NT application. Liquid junction potentials (estimated using the Clampex software) were corrected offline.

Statistical analyses. Analyses were performed using Clampfit, Sigma Plot and Origin. Data are presented as mean ± s.e.m.

30. Liu, J., Xia, J., Cho, K. H., Clapham, D. E. & Ren, D. Catsper β: a novel transmembrane protein in the catsper channel complex. *J. Biol. Chem.* **282**, 18945–18952 (2007).
31. Masuko, S., Nakajima, S. & Nakajima, Y. Dissociated high-purity dopaminergic neuron cultures from the substantia nigra and the ventral tegmental area of the postnatal rat. *Neuroscience* **49**, 347–364 (1992).
32. Rayport, S. *et al.* Identified postnatal mesolimbic dopamine neurons in culture: morphology and electrophysiology. *J. Neurosci.* **12**, 4264–4280 (1992).
33. Grace, A. A. & Onn, S.-P. Morphology and electrophysiological properties of immunocytochemically identified rat dopamine neurons recorded in vitro. *J. Neurosci.* **9**, 3463–3481 (1989).

LETTERS

Counting RAD51 proteins disassembling from nucleoprotein filaments under tension

Joost van Mameren[1]†, Mauro Modesti[2,3], Roland Kanaar[2,4], Claire Wyman[2,4], Erwin J. G. Peterman[1]* & Gijs J. L. Wuite[1]*

The central catalyst in eukaryotic ATP-dependent homologous recombination consists of RAD51 proteins, polymerized around single-stranded DNA. This nucleoprotein filament recognizes and invades a homologous duplex DNA segment[1,2]. After strand exchange, the nucleoprotein filament should disassemble so that the recombination process can be completed[3]. The molecular mechanism of RAD51 filament disassembly is poorly understood. Here we show, by combining optical tweezers with single-molecule fluorescence microscopy and microfluidics[4,5], that disassembly of human RAD51 nucleoprotein filaments results from the interplay between ATP hydrolysis and the release of the tension stored in the filament. By applying external tension to the DNA, we found that disassembly slows down and can even be stalled. We quantified the fluorescence of RAD51 patches and found that disassembly occurs in bursts interspersed by long pauses. After relaxation of a stalled complex, pauses were suppressed resulting in a large burst. These results indicate that tension-dependent disassembly takes place only from filament ends, after tension-independent ATP hydrolysis. This integrative single-molecule approach allowed us to dissect the mechanism of this principal homologous recombination reaction step, which in turn clarifies how disassembly can be influenced by accessory proteins.

Homologous recombination is a vital mechanism that maintains genome integrity by repairing double-strand breaks in DNA, and generates genetic diversity by exchanging DNA between chromosomes during meiosis. The central process in homologous recombination is the strand exchange between homologous DNA segments. Recombinase proteins such as RecA and RAD51 catalyse this process by forming an ATP-dependent helical filament around single-stranded DNA (ssDNA)[1]. This filament finds a homologous segment of double-stranded DNA (dsDNA), invades it and catalyses strand exchange to generate a joint molecule. This resulting structure is further processed in several steps by additional proteins, finally yielding two intact, homologous dsDNAs[1,2]. For these steps to proceed properly, it is essential that RAD51 filaments disassemble. Hydrolysis of ATP bound at the interface between adjacent monomers is a prerequisite for filament disassembly[6–8]. RecA and RAD51 not only form ATP-dependent filaments on ssDNA but also on dsDNA[4,9–11]. These dsDNA nucleoprotein filaments may have deleterious effects *in vivo*, for example by sequestering these recombinases in nonfunctional complexes that could obstruct other DNA transactions. RAD51 filament disassembly can be aided by auxiliary proteins[3]. To understand recombinase removal, it is necessary to determine the molecular mechanism of the intrinsic RAD51 disassembly reaction.

To follow this process under controlled conditions, we developed an instrument that combines fluorescence microscopy with force-measuring dual optical traps[4] and a custom-built multichannel microfluidic flow cell (Supplementary Fig. 1)[11–13]. This instrument enabled us to control and trigger biochemical reactions at the same time as mechanically manipulating individual DNA molecules. It also allowed us to image and quantify the fluorescence from functional human RAD51 variants with a single surface-exposed cysteine, labelled with Alexa Fluor 555 (refs 4 and 14).

Our experimental assay, in which we moved single Ca^{2+}-stabilized RAD51–dsDNA complexes[4,15,16] to a Mg^{2+}-containing buffer by swiftly shifting the microscope stage between parallel flow channels, is depicted in Fig. 1a. This buffer exchange activates ATP hydrolysis. Figure 1b shows a kymograph[4] of fluorescently labelled RAD51 polymerized onto a dsDNA molecule, held from one side by a single optically trapped bead and stretched by buffer flow (Supplementary Video 1). The triggered ATP hydrolysis results in filament disassembly, evidenced by a steady decrease of intensity and a marked shrinkage of the complex. This shrinkage, caused by relief of RAD51-induced DNA extension, immediately excludes photobleaching as the cause of intensity decrease[4]. Some patches seem to shrink from their ends (for example, the one marked with an asterisk, Fig. 1b), suggesting that disassembly occurs from filament ends, as reported for RecA[17–19]. Using a more sophisticated analysis, we will address this question in more detail later.

RAD51 forms helical filaments on dsDNA that extend the DNA by about 50% compared to B-form DNA[4,7,10,14]. It is possible that the tension thus stored forms a driving force for the disassembly process[20]. To test this hypothesis, we captured RAD51–DNA complexes from both ends between two optically trapped beads[4,13]. Figure 2a shows the time course of fluorescence intensity and tension for a complex undergoing disassembly while being held at fixed end-to-end distance (Supplementary Video 2). Owing to the shrinking contour length, the DNA pulls itself taut, after which tension gradually builds up. We observed that disassembly slowed down with increasing tension and even stalled at a tension of 48 ± 3 pN (s.e.m.; $n = 7$). To test that this slowing down is an actual characteristic of RAD51 and not due to the decreasing number of monomers left to dissociate, we examined the effect of a sudden tension release, induced by instantaneously moving the optical traps closer together. Indeed, disassembly immediately reinitiates after tension release (Fig. 2b), confirming the stabilizing effect of tension on RAD51 filaments.

The disassembly rate, calculated as the time derivative of the fluorescence intensity, decreased with tension (Fig. 2c). Apparently, the energy barrier of disassembly is raised by a tension increase, stabilizing the RAD51-bound state. The rate decrease is well fit by a single exponential, suggesting a dependence according to Arrhenius's law: $k(F) = k(0)\exp[-x^{\dagger}F/k_{\mathrm{B}}T]$, in which $k_{\mathrm{B}}T$ is the thermal energy, F is

[1]Laser Centre and Department of Physics and Astronomy, VU University, De Boelelaan 1081, 1081 HV, Amsterdam, The Netherlands. [2]Department of Cell Biology and Genetics, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, The Netherlands. [3]CNRS, Unité Propre de Recherche 3081, Genome Instability and Carcinogenesis Conventionné par l'Université d'Aix-Marseille 2, 13402 Marseille Cedex 20, France. [4]Department of Radiation Oncology, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, The Netherlands. †Present address: JPK Instruments AG, Bouchéstrasse 12, 12435 Berlin, Germany.
*These authors contributed equally to this work.

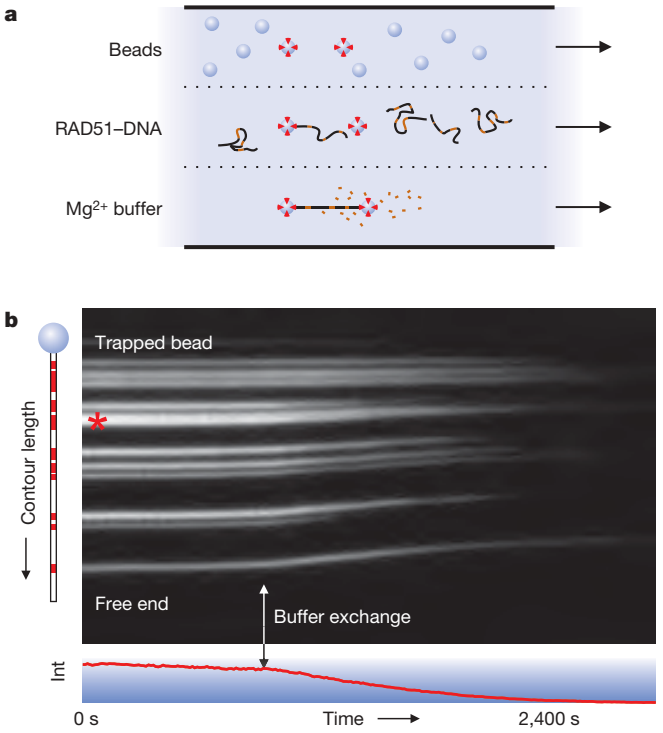


Figure 1 | Assay for triggering of RAD51 disassembly. **a**, Schematic of the multichannel flow cell. After capturing RAD51-bound dsDNA molecules by one or both ends using optical traps, ATP hydrolysis is triggered by moving the complex to a Mg^{2+}-containing channel, setting off disassembly. **b**, Kymograph of a RAD51-coated dsDNA molecule held from one end, stretched by flow. RAD51 patches are marked red in the cartoon on the left. Mg^{2+}-induced filament disassembly (vertical arrow) is evidenced by simultaneous DNA contraction and a steady decrease in fluorescence intensity (int; bottom graph).

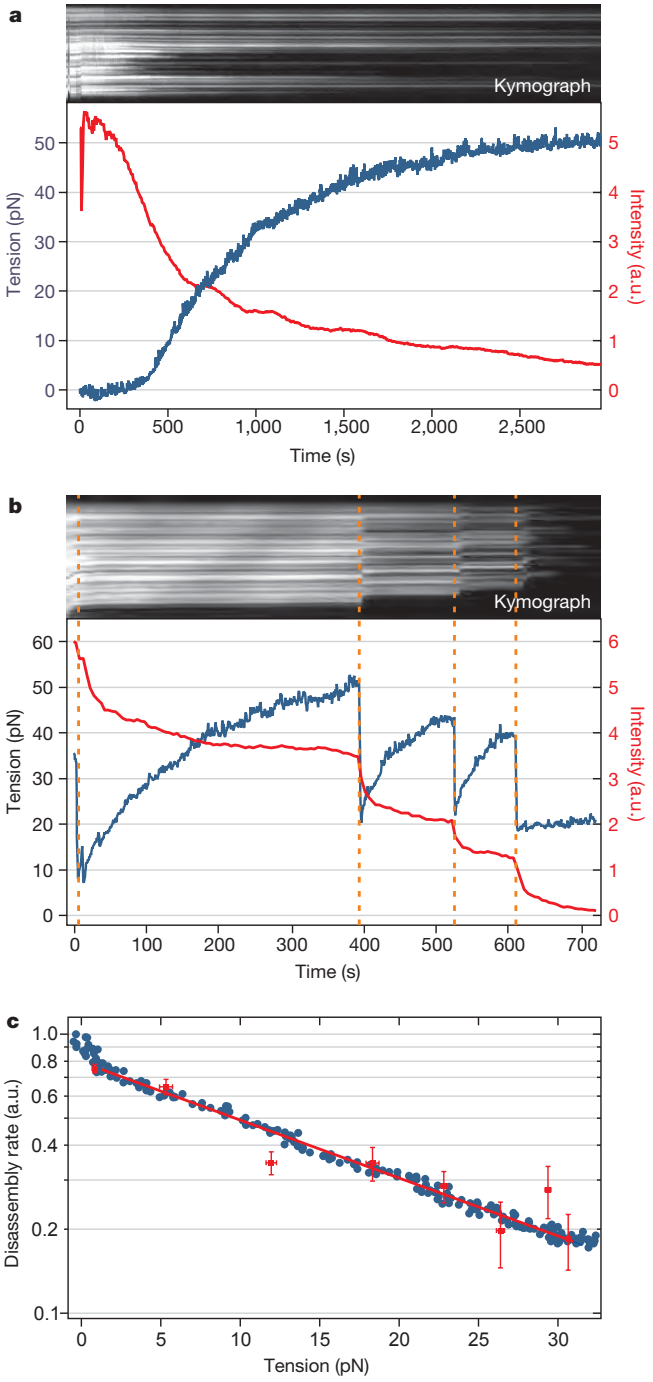


Figure 2 | RAD51 disassembly rate is reversibly reduced by DNA tension. **a**, Kymograph and intensity trace (red) of a RAD51–dsDNA complex, held at fixed length, triggered to disassemble at $t = 0$. Tension (blue) increases owing to disassembly-induced DNA contraction, levelling off to stall at around 50 pN. The intensity decrease slows down accordingly. **b**, Tension-stalled disassembly is reinitiated by tension release (orange dashes). **c**, Disassembly rates decrease exponentially with tension. Directly differentiated intensity trace (red symbols) and smoothed trace (blue) are fitted by Arrhenius's equation, yielding the same $x^{\dagger}$ value (0.20 ± 0.01 nm for this complex). See Supplementary Information for calculation.

the tension, and $x^{\dagger}$ denotes the distance to the transition state along the relevant reaction coordinate[21]. This transition state is intermediate between the initial state with RAD51 bound to extended DNA, and the final one with relaxed DNA without RAD51. We determined $x^{\dagger}$ to be 0.27 ± 0.04 nm (s.e.m., $n = 9$). One RAD51 monomer covers three base pairs and holds them in an extended conformation of about 1.5 nm in length (compared to 1 nm in canonical B-form DNA)[4,7,10,16]. Therefore, on disassembly of a single monomer the DNA shrinks at most by half a nanometre. Hence our value for the location of the transition state, $x^{\dagger}$, is consistent with filaments disassembling one monomer at a time.

We next sought to demonstrate more directly that RAD51 filaments disassemble as monomers from filament ends, and to extract kinetic rates. To address these questions, we calibrated the fluorescence intensity to numbers of RAD51 monomers using single-molecule photobleaching steps in Ca^{2+}-stabilized, optically trapped RAD51–DNA complexes (Supplementary Figs 2 and 3). Figure 3a shows a kymograph and corresponding disassembly traces of four isolated RAD51 patches. A notable feature emerges: the intensity decrease is not continuous, but occurs in bursts of varying size, interspersed with pauses in the order of minutes. We fitted many such isolated disassembling filaments using a step-fitting algorithm (blue lines in Fig. 3a and see Supplementary Fig. 2)[22]. Using the fitted steps, the kinetics of the disassembly can be analysed from distributions of pause durations and burst sizes (Fig. 3b and Supplementary Figs 4 and 6). Pause durations are exponentially distributed with a time constant of 152 ± 9 s, suggesting that the pauses were caused by a single Poisson waiting step in the reaction.

This burst-wise disassembly can be understood by a model in which monomers dissociate exclusively from filament ends after ATP hydrolysis (graphically depicted in Fig. 4a). Assuming ATP hydrolysis takes place uniformly along the nucleoprotein filament (such as RecA[23]), we interpret pauses as events in which filament disassembly transiently halts because the terminal monomer has ATP bound. Once that ATP is hydrolysed, the terminal monomer loses contact with the DNA and dissociates, as do the neighbours that have already hydrolysed their ATP. This burst of disassembly stops once an ATP-bound monomer is encountered. An inference of this

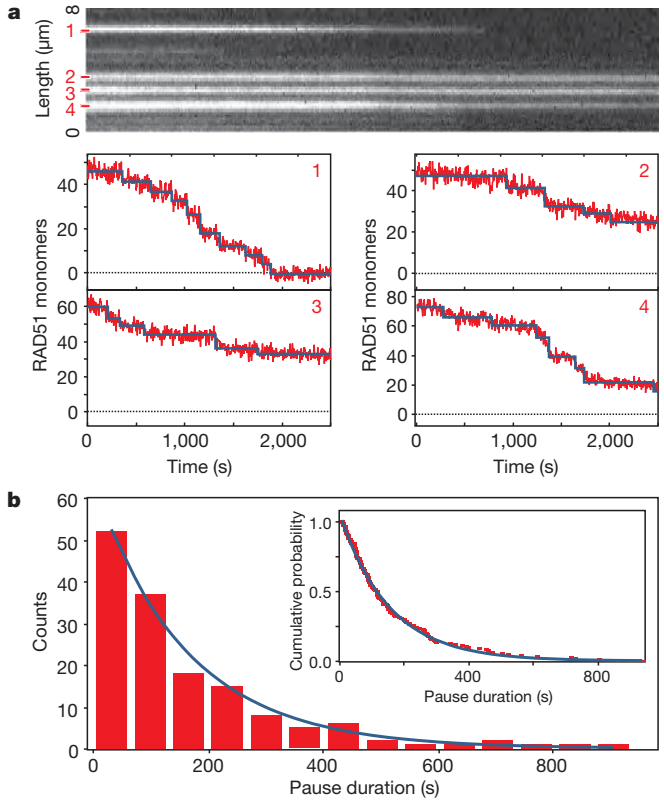


Figure 3 | RAD51 disassembly occurs in bursts interspersed with pauses. **a**, Calibrated intensity traces of isolated, short RAD51 patches show a burst pattern of disassembly activity interspersed by pauses in the order of minutes. These staircases are well fitted by a step fitting routine (blue lines)[22]. **b**, Pause durations are exponentially distributed, both when binned into a histogram (decay constant 152 ± 9 s) and when plotted as a cumulative probability distribution (inset, decay constant 152 ± 1 s).

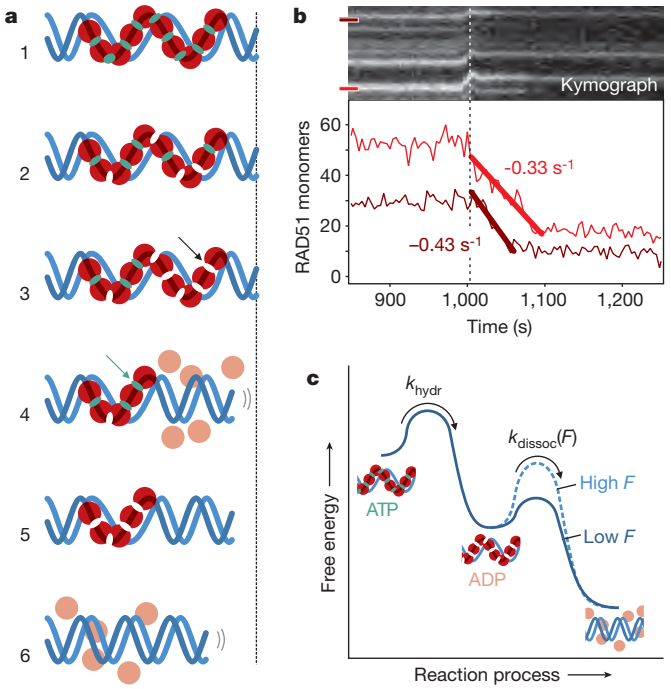


Figure 4 | RAD51 disassembly pathway. **a**, (1) All RAD51 (red) starts out ATP- (green) and DNA- (blue) bound. (2) ATP hydrolysis is triggered; filaments remain stable as long as terminal RAD51s have ATP bound. (3) ATP hydrolysis at terminal monomer. (4) Monomers dissociate until next ATP-bound monomer is terminal (arrow). (5) Disassembly pauses until terminal ATP hydrolyses. (6) Disassembly relaxes DNA. **b**, Stalled disassembly is reinitiated by tension release (dotted line). The slopes of the prolonged bursts yield the intrinsic dissociation rate. **c**, Free energy diagram of RAD51 disassembly. After ATP hydrolysis by a terminal monomer (with tension-independent rate k_{hydr}), ADP-bound monomers dissociate with rate $k_{dissoc}(F)$ that exponentially decreases with tension. Tension increases either the energy of the DNA-bound ADP-state, or the energy barrier to the dissociated state (as depicted).

model is that the ATPase rate of RAD51 bound to dsDNA is the reciprocal of the average pause duration. To determine this rate accurately we need to take into account the fact that the fit residuals in the pause plateaus show a small but non-zero average slope (-0.01 monomers s^{-1}). This indicates that small steps (1–3 monomers) are hidden in the noise in our single-patch intensity traces (Fig. 3a) under the applied illumination conditions. From this we could determine that on average one short disassembly event per fitted pause was not detected, and that we thus overestimated the average pause duration by a factor of 2. Taking this into account, we determined that the ATP hydrolysis rate, k_{hydr}, is $0.6–1.3 \times 10^{-2}\,s^{-1}$. This value is similar to that measured with bulk chemical kinetics assays[24], which confirms that the observed pauses are governed by ATP hydrolysis. Interestingly, burst-wise disassembly was still observed in the presence of 2 mM ATP ($k_{hydr} = 1.0 \times 10^{-2}\,s^{-1}$; Supplementary Fig. 4). This suggests that ATP renewal along the filaments takes place at a considerably slower rate than hydrolysis, if occurring at all, suggesting that ADP release is slower than disassembly. This is consistent with ATPase assays that showed that ADP release is the rate-limiting step in the ATPase cycle[24–26]. Our model also predicts the shape of the intensity curves such as that shown in Fig. 2a (see Supplementary Fig. 5), yielding an estimate for the average filament length of 10–50 monomers, in agreement with previous results[16]. Moreover, we can predict that on average 5–10 monomers are involved in a burst, which is in agreement with our measurements (compare the total numbers of bursts in Fig. 3a and Supplementary Fig. 6).

In our disassembly model, ATP hydrolysis precedes dissociation of monomers. A question remains as to which of these two causes the reaction to stall when tension is applied to the DNA (Fig. 2). In case only dissociation would depend on tension, ATP hydrolysis would proceed even when disassembly is stalled with high tension. When such a stalled complex is relaxed, an extended disassembly burst without pauses is expected. Indeed, isolated stalled patches exhibit such extended disassembly (Fig. 4b), confirming that ATP hydrolysis continues even at high DNA tension. Apparently, DNA tension changes which reaction step is rate-limiting: ATP hydrolysis on filament ends at no tension and monomer dissociation at high tension. The extended bursts in Fig. 4b allowed us to directly determine the intrinsic dissociation rate of RAD51 monomers at low tension from the slope of the intensity decrease: $k_{dissoc} = 0.51 \pm 0.14\,s^{-1}$ (s.e.m., $n = 5$). The complete mechanokinetic model for ATP hydrolysis and tension-dependent RAD51 disassembly is summarized in a free-energy diagram in Fig. 4c. The extension of DNA that RAD51 imposes acts as a loaded spring that, in part, drives the disassembly reaction. For this to occur, the RAD51–DNA complex must shorten by $x^{\dagger}$ to reach the transition state through thermal fluctuations. The amplitudes of thermal fluctuations shortening the DNA are reduced by external tension, kinetically disfavouring disassembly. In contrast, ATP hydrolysis seems to be independent of DNA tension, suggesting that hydrolysis does not alter the extended conformation of the DNA.

Our data and model illustrate how RAD51 filament stability crucially depends on the nucleotide state of the filament terminus. When the terminal RAD51 is in the ATP-bound state, the filament end is stably attached to the DNA. After hydrolysis, the terminal RAD51 loses its DNA affinity and detaches. Within a filament, RAD51 monomers seem to be locked onto the DNA by their neighbours, independent of the nucleotide state. In RecA–DNA co-crystals, such coupling between nucleotide state, monomer–monomer interactions and

recombinase binding to DNA has also been observed[27]. Notably, with our data and model it now becomes possible to shed new light on the way accessory proteins can catalyse filament disassembly. First, our data show that RAD51 nucleoprotein filament ends dominate the disassembly process, clarifying why proteins such as RAD54 interact with filament termini to stimulate disassembly[28]. If RAD54 destabilizes the interaction between the terminal RAD51 and DNA, this would result in accelerated disassembly by decreasing the duration of pauses that we show occur during unaided RAD51 disassembly. Second, it explains why stable RAD51 filaments in the presence of ADP can only form when there is at least some ATP present[29]: the end caps need to contain ATP to stabilize a filament. This, in turn, clarifies why the rate of RAD54-assisted filament disassembly is higher in the presence of ADP[30]. These ADP-containing filaments would give rise to longer disassembly bursts and RAD54 would need to remove ATP-stabilized end caps less often. Thus, concurrent visualization, quantification, manipulation and triggering of the dynamics of RAD51 filaments provided a comprehensive understanding of spontaneous RAD51 nucleoprotein filament disassembly, which in turn allowed new insights in the disassembly assisted by accessory proteins. We expect that integrative single-molecule approaches, such as those used here, will also be of great value in dissecting many other complex biological reactions.

METHODS SUMMARY

Biotinylated λ-DNA and fluorescently labelled human RAD51 (isoform Q313) were prepared as described elsewhere[12–14]. RAD51–dsDNA nucleoproteins were pre-assembled in Ca^{2+}-stabilized conditions as used before[4,14]. ATP hydrolysis was triggered in the flow cell by exposing the trapped RAD51–dsDNA complex to a buffer containing 10 mM Mg^{2+} and 10 mM EGTA.

Descriptions of the combined dual optical tweezers and fluorescence microscope[4] as well as data analysis procedures are provided in the Supplementary Information and Supplementary Figs 1–3.

Received 30 October 2007; accepted 24 October 2008.
Published online 7 December 2008.

1. Bianco, P. R., Tracy, R. B. & Kowalczykowski, S. C. DNA strand exchange proteins: a biochemical and physical comparison. *Front. Biosci.* **3,** D570–D603 (1998).
2. Sung, P., Krejci, L., Van Komen, S. & Sehorn, M. G. Rad51 Recombinase and Recombination Mediators. *J. Biol. Chem.* **278,** 42729–42732 (2003).
3. Symington, L. S. & Heyer, W. D. Some disassembly required: role of DNA translocases in the disruption of recombination intermediates and dead-end complexes. *Genes Dev.* **20,** 2479–2486 (2006).
4. van Mameren, J. *et al.* Dissecting elastic heterogeneity along DNA molecules coated partly with Rad51 using concurrent fluorescence microscopy and optical tweezers. *Biophys. J.* **91,** L78–L80 (2006).
5. Brau, R. R. *et al.* Interlaced optical force-fluorescence measurements for single molecule biophysics. *Biophys. J.* **91,** 1069–1077 (2006).
6. Kowalczykowski, S. C. & Eggleston, A. K. Homologous pairing and DNA strand-exchange proteins. *Annu. Rev. Biochem.* **63,** 991–1043 (1994).
7. Benson, F. E., Stasiak, A. & West, S. C. Purification and characterization of the human Rad51 protein, an analogue of *E. coli* RecA. *EMBO J.* **13,** 5764–5771 (1994).
8. Chi, P. *et al.* Roles of ATP binding and ATP hydrolysis in human Rad51 recombinase function. *DNA Repair (Amst.)* **5,** 381–391 (2006).
9. Hegner, M., Smith, S. B. & Bustamante, C. Polymerization and mechanical properties of single RecA–DNA filaments. *Proc. Natl Acad. Sci. USA* **96,** 10109–10114 (1999).
10. Ristic, D. *et al.* Human Rad51 filaments on double- and single-stranded DNA: correlating regular and irregular forms with recombination function. *Nucleic Acids Res.* **33,** 3292–3302 (2005).
11. Galletto, R., Amitani, I., Baskin, R. J. & Kowalczykowski, S. C. Direct observation of individual RecA filaments assembling on single DNA molecules. *Nature* **443,** 875–878 (2006).
12. Noom, M. C., van den Broek, B., van Mameren, J. & Wuite, G. J. L. Visualizing single DNA-bound proteins using DNA as a scanning probe. *Nat. Methods* **4,** 1031–1036 (2007).
13. van den Broek, B., Noom, M. C. & Wuite, G. J. DNA-tension dependence of restriction enzyme activity reveals mechanochemical properties of the reaction pathway. *Nucleic Acids Res.* **33,** 2676–2684 (2005).
14. Modesti, M. *et al.* Fluorescent human RAD51 reveals multiple nucleation sites and filament segments tightly associated along a single DNA molecule. *Structure* **15,** 599–609 (2007).
15. Bugreev, D. V. & Mazin, A. V. Ca^{2+} activates human homologous recombination protein Rad51 by modulating its ATPase activity. *Proc. Natl Acad. Sci. USA* **101,** 9988–9993 (2004).
16. van der Heijden, T. *et al.* Real-time assembly and disassembly of human RAD51 filaments on individual DNA molecules. *Nucleic Acids Res.* **35,** 5646–5657 (2007).
17. Lindsley, J. E. & Cox, M. M. Assembly and disassembly of RecA protein filaments occur at opposite filament ends. Relationship to DNA strand exchange. *J. Biol. Chem.* **265,** 9043–9054 (1990).
18. Arenson, T. A., Tsodikov, O. V. & Cox, M. M. Quantitative analysis of the kinetics of end-dependent disassembly of RecA filaments from ssDNA. *J. Mol. Biol.* **288,** 391–401 (1999).
19. Joo, C. *et al.* Real-time observation of RecA filament dynamics with single monomer resolution. *Cell* **126,** 515–527 (2006).
20. Wyman, C. Monomer networking activates recombinases. *Structure* **14,** 949–950 (2006).
21. Evans, E. Probing the relation between force—lifetime—and chemistry in single molecular bonds. *Annu. Rev. Biophys. Biomol. Struct.* **30,** 105–128 (2001).
22. Kerssemakers, J. W. J. *et al.* Assembly dynamics of microtubules at molecular resolution. *Nature* **442,** 709–712 (2006).
23. Brenner, S. L. *et al.* RecA protein-promoted ATP hydrolysis occurs throughout recA nucleoprotein filaments. *J. Biol. Chem.* **262,** 4011–4016 (1987).
24. Tombline, G. & Fishel, R. Biochemical characterization of the human RAD51 protein. I. ATP hydrolysis. *J. Biol. Chem.* **277,** 14417–14425 (2002).
25. Tombline, G., Shim, K. S. & Fishel, R. Biochemical characterization of the human RAD51 protein. II. Adenosine nucleotide binding and competition. *J. Biol. Chem.* **277,** 14426–14433 (2002).
26. Shim, K. S. *et al.* Magnesium influences the discrimination and release of ADP by human RAD51. *DNA Repair (Amst.)* **5,** 704–717 (2006).
27. Chen, Z., Yang, H. & Pavletich, N. P. Mechanism of homologous recombination from the RecA–ssDNA/dsDNA structures. *Nature* **453,** 489–494 (2008).
28. Kiianitsa, K., Solinger, J. A. & Heyer, W. D. Terminal association of Rad54 protein with the Rad51–dsDNA filament. *Proc. Natl Acad. Sci. USA* **103,** 9767–9772 (2006).
29. Zaitseva, E. M., Zaitsev, E. N. & Kowalczykowski, S. C. The DNA binding properties of *Saccharomyces cerevisiae* Rad51 protein. *J. Biol. Chem.* **274,** 2907–2915 (1999).
30. Li, X. *et al.* Rad51 and Rad54 ATPase activities are both required to modulate Rad51–dsDNA filament dynamics. *Nucleic Acids Res.* **35,** 4124–4140 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. van den Broek and R.T. Dame for discussions and a critical reading of the manuscript and J. Kerssemakers for kindly providing his step fitting algorithm. This work was supported by the Biomolecular Physics program of the Dutch organization for Fundamental Research of Matter (FOM) (to R.K., C.W., E.J.G.P. and G.J.L.W.), and grants from the Dutch Cancer Society (KWF), the Netherlands Organization for Scientific Research (NWO), the Netherlands Genomics Initiative/NWO, the Association for International Cancer Research and the European Commission Integrated Projects Molecular Imaging and DNA Repair and a National Cancer Institute–National Institutes of Health USA program project (to C.W. and R.K.). E.J.G.P. and G.J.L.W. are recipients of NWO Vidi grants; C.W. of an NWO Vici grant.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.J.G.P (erwinp@few.vu.nl) or G.J.L.W. (gwuite@few.vu.nl).

naturejobs

THE CAREERS MAGAZINE FOR SCIENTISTS

The international economic downturn could have a curious by-product: more demand for top scientists — at least in the short term. In the United States, lawmakers are creeping closer to a stimulus package that would provide billions of dollars for the National Institutes of Health and the National Science Foundation (see *Nature* **457,** 623; 2009). And in China, the government is offering top Chinese professors who are working overseas relocation packages of 1 million renminbi (US$146,000) per person to lure them back to the mainland (see *Nature* **457,** 522; 2009). With college-educated Chinese students struggling with an increasingly tough job market, the timing of this initiative may be more than just a coincidence.

Governments are now looking to scientists to help them build solid investments for the future. A top-notch science workforce is viewed as a reliable path to innovation, economic growth and cutting-edge industries. In addition to the stimulus package, US President Barack Obama has touted aspirations to create 'green jobs'— setting aside billions of dollars a year for the next decade or so to invest in renewable energy. The idea is to create five million green jobs that have good salaries, can't be outsourced and will help to end the nation's dependence on foreign oil, says Obama.

As with many large, bulk investments, not all the money set aside will directly benefit scientists, and some researchers will benefit more than others (see page 750). In the United States, for instance, much of the funds will go towards infrastructure. And irrespective of how big they are, short-term infusions of cash do not always translate into sustained, decades-long successes in science, or even into sustainable budgets at science agencies.

But perhaps a window of opportunity is opening up. Some researchers will be able to benefit from governments seeking the best and the brightest science and engineering talent to help boost sluggish economies. And if the funds come through, scientists, especially young researchers, may have a better — albeit fleeting — chance to earn grants and establish themselves and their careers.

Gene Russo is editor of *Naturejobs*.

CONTACTS
Editor: Gene Russo
Assistant editor: Karen Kaplan
e-mail: naturejobseditor@naturedc.com

European Head Office, London
The Macmillan Building,
4 Crinan Street, London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Dan Churchward (4966)
e-mail: d.churchward@nature.com
Assistant European Manager:
Nils Moeller (4953)
***Nature*events:**
Ghizlaine Ababou (+44 (0) 20 7014 4015)
e-mail: g.ababou@nature.com
Southwest UK/RoW:
Alexander Ranken (4944)

Northeast UK/Ireland:
Matthew Ward (+44 (0) 20 7014 4059)
France/Switzerland/Belgium:
Muriel Lestringuez (4994)
Scandinavia/Spain/Portugal/Italy:
Evelina Rubio-Hakansson (4973)
North Germany/The Netherlands/Eastern Europe: Kerstin Vincze (4970)
South Germany/Austria:
Hildi Rowland (+44 (0) 20 7014 4084)
Advertising Production Manager:
Stephen Russell
To send materials use London address above.
Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com
***Naturejobs* web development:** Tom Hancock
***Naturejobs* online production:** Dennis Chu

US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com
US Sales Manager: Ken Finnegan

India
Vikas Chawla (+91 1242881057)
e-mail: v.chawla@nature.com

Japan Head Office, Tokyo
Chiyoda Building, 2-37 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe (+81 3 3267 8765)
e-mail: a.watanabe@natureasia.com
Business Development Manager, Greater China/Singapore:
Gloria To (+852 2811 7191)
e-mail: g.to@natureasia.com

K. REDINGER/DESIGN PICS/CORBIS

SALARIES IN THE BALANCE

Postdoc salaries vary widely at every level, from countries down to individual teams. **Paul Smaglik** looks at where the problems lie.

When Yegor Domanov moved from Ukraine to the University of Helsinki on a fellowship funded by the Academy of Finland four years ago, he received a monthly stipend of €2,000 (US$2,600). The good news was that he didn't have to pay any taxes on it. The bad news was that he received no benefits, so he opted not to pay for health-care coverage. Now, he is in the second year of a €4,500-a-month Marie Curie fellowship administered by the European Commission. However, because taxes, pension and health care are deducted from his stipend, his take-home base salary is €2,200 a month. But he also receives a non-taxable monthly 'mobility supplement' — an incentive to encourage researchers to ferry knowledge to different parts of Europe — that adds another €600 a month to his pay.

Domanov's next fellowship, at the Curie Institute in Paris and funded by the city's Fondation Pierre-Gilles de Gennes, will cover health insurance and social-security deductions and include pension contributions and some local-transportation costs. But the €2,800 base will be taxable. Although French taxes aren't as high as Finland's, Domanov expects to earn about the same base salary he now receives. Still, he is holding out hope that he can actually grow his pay along with his career. He is applying to the European Commission for a two-year, €15,000-per-year 're-entry' grant intended to help Marie Curie fellows establish themselves in another country.

Although postdoc stipends start at a base of about US$40,000 in the United States and €30,000 in Europe, what fellows can expect to take home varies between countries, funding agencies and even grant types within an agency. Differing policies on taxation, health benefits, pensions and supplements can all combine to make calculating the take-home pay for potential fellowships a complex prospect.

V. TRUSOVA.

Optimist: Yegor Domanov.

Foundations, funding agencies and individual institutions are all working to create a fairer compensation scheme for postdocs. But even if the base stipends seem the same, prospective postdocs should investigate the variables before signing on, otherwise they might be unpleasantly surprised by their take-home pay.

Solid guidelines

Domanov credits the European Commission's research charter for improving postdoc benefits in Europe. Detlev Arendt, head of postdoctoral training at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany, says that a growing number of European postdocs are getting European Commission grants, which include provisions for mobility and benefits such as health care. Even though EMBL's postdocs aren't funded through the commission, the institution has tried to stay competitive by adopting many of the recommendations set out by the charter, which EMBL signed last year (see *Nature* **455,** 426–428; 2008).

For example, EMBL offers supplements of €300 a month for each dependent — including children and unemployed spouses. It is also now preparing a pension scheme that will allow fellows to contribute towards their retirement. The institution looks to a host of European funders — such as the Alexander von Humboldt Foundation; Germany's main research funding agency, the DFG; and Britain's Wellcome Trust

— to set its starting base salary of €2,500 a month.

Around the world, various groups are trying to push the base higher. This is especially important because a major benchmark, the National Research Service Award provided by the US National Institutes of Health (NIH), has remained flat for the past three years at about $3,083 (€2,300) per month for new postdocs (see 'Follow the leader'). The scale increases incrementally for those with more experience, topping out at $4,250 a month for postdocs with seven or more years of experience.

Foundations have a history of offering better salaries and benefits to postdocs than government funding agencies. And in doing so, they aim to prod the government to up their contributions accordingly. The Wellcome Trust, for example, offers supplements to 'top up' stipends, thus making these positions more desirable than those funded, say, by the UK Medical Research Council. The trust adds £2,500 (US$3,510) a year to recipients of training and junior fellowships, £7,500 a year for intermediate fellowships and £12,500 a year for senior fellowships.

Anthony Woods, head of medicine, society and history grants at the Wellcome Trust, says that these enhancements are necessary to keep top people in science — and in Britain. "We do not want these people being lost from science because of low salaries," Woods says. "We want to make science an attractive career."

Foundations such as the Wellcome Trust and the Howard Hughes Medical Institute in Chevy Chase, Maryland, have conventionally tried to pay higher salaries and prompt the government to match them. "We were trying to force the government agencies to pay higher," Woods says.

At Howard Hughes, that means paying attention to baseline levels such as the National Research Service Award amount, then consistently topping them. Even though the principal investigators at Howard Hughes set the salaries for their fellows, only about 10% of the fellows receive the baseline rate; the remainder are paid at higher levels, says Phil Perlman, a grants officer at the institute. Howard Hughes' stipends are taxable, and the institute doesn't allow fellows to supplement their salaries with other grants, but it does provide additional money to pay for health care if the fellow's host organization doesn't provide it, he says.

Grant-giver: Phil Perlman.

UT SOUTHWESTERN MEDICAL CENTER AT DALLAS

Leaps and lulls

NIH officials are aware of this trend. Walter Schaffer, senior scientific adviser of the agency's Office of Extramural Research, says that in the past 25 years, NIH stipends have been "spotty". "There have been big jumps followed by long periods of quiescence," he says.

Maxine Singer, author of the National Academies of Science report *Enhancing the Postdoctoral Experience for Scientists and Engineers* and emeritus president of the Carnegie Institution in Washington DC, says that addressing stipends alone won't improve the plight of postdocs. Although postdoc salaries have been generally too low and flat for too long, she says, bigger issues are the growing length and number of fellowships young scientists now face. "When people did their PhDs in four years and a postdoc for two years, low stipends weren't such a big deal," Singer says. In the life sciences, the time it takes to earn a PhD has now grown to an average of seven years and the length of an average postdoc has nearly doubled to four.

Domanov says that finding generous fellowship programmes requires more effort and means not only researching career-advancement schemes but also being aware of issues such as potential taxes, top-ups and benefits. He advises fellows to begin their search for such programmes at least a year before their existing fellowship expires. Domanov, who is secretary-general of the Marie Curie Fellows Association, says that the association can help postdocs through the confusion. It provides useful information on salaries, allowances, taxes and pensions to postdocs in Europe and European postdocs working abroad — not just to Marie Curie fellows. "There are opportunities if one is willing to spend some time in research for funding." ■

Paul Smaglik is moderator of the Nature Network career site.

FOLLOW THE LEADER

The US National Postdoctoral Association (NPA) is advocating for higher stipends from the National Research Service Award (NRSA) given to postdocs by the National Institutes of Health (NIH). Many funders around the world use the NRSA as a baseline to determine their own annual stipend levels. However, policy watchers are making cases both for and against bumping up this baseline, which is currently US$37,000 for new postdocs.

The last major change to the NRSA came when the NIH was in the middle of a cycle that awarded more and larger grants to principal investigators but left some postdocs behind. In 2000, the US National Academy of Sciences recommended that postdoc stipends be increased by 3% a year to keep up with inflation and the cost of living. "We responded with gradual increases," says Walter Schaffer, senior scientific adviser of extramural research at the NIH. "But since then, they have been relatively flat."

Schaffer says that increasing stipends would incur trade-offs that many would find unsatisfactory. Budget constraints — the $29-billion NIH budget has remained relatively flat since 2005 — mean that there would be either more postdoc positions with low stipends or fewer fellowships with higher compensation.

Last year, the NPA started to lobby the US Congress for higher stipends. The association's executive director, Cathee Johnson-Phillips, has since advocated for a bigger overall NIH budget, which looks likely to happen under US President Barack Obama's economic stimulus plan (see *Nature* **457,** 364–365; 2009), in the hope that it will translate into higher postdoc salaries. However, Maxine Singer, author of a 2000 National Academies of Science report, entitled *Enhancing the Postdoc Experience for Scientists and Engineers*, and president emeritus of the Carnegie Institution in Washington DC, warns that large budget rises don't necessarily translate into higher stipends for all postdocs — some receive more of the pot than others. She says that conditions need to improve for postdocs in the life sciences particularly — they should be receiving higher stipends and spending less time in their role. "In other fields, stipends for postdocs have been higher," she explains.

Phil Perlman, a grants officer at the Howard Hughes Medical Institute in Chevy Chase, Maryland, says that using the NRSA as a benchmark is a relatively new phenomenon. "There was a time when the NIH was not the benchmark," Perlman says. "The NRSA stipends were well below the market. If you got an NRSA, you took a salary cut."

Perlman says that he isn't sure that now is the right time to boost the NRSA level. At present, the stipends include mandatory raises "that are quite large and don't have any evaluation of performance". He thinks that before it revisits the baseline stipend level, the NIH should implement some performance standards and stop handing out automatic raises. And other funders around the world should look to market forces, not to the NRSA, to establish their own stipend levels, he says. "Right now, I don't know how much of an economic case they can make for increasing their baseline," Perlman says. **P.S.**

NATURE|Vol 457|5 February 2009

MOVERS

Thomas Henzinger, president, Institute of Science and Technology Austria, Klosterneuburg, Austria

2004–09: Professor, computer and communication sciences, Swiss Federal Institute of Technology, Lausanne, Switzerland
1998–2009: Professor and adjunct professor (since 2005), electrical engineering and computer sciences, University of California, Berkeley

As president of the Institute of Science and Technology (IST) Austria, Europe's newest science and technology academy in Klosterneuburg, Tom Henzinger's focus will shift from researching ways to improve the reliability of software and hardware systems to developing a world-class institute, faculty and staff.

Henzinger, a renowned computer scientist who is best known for his work in real-time and embedded systems, formally assumes his new post on 1 September. IST Austria's buildings and campus are still under construction, and its newness was one of the reasons he was drawn to the position. "That is what attracted me," Henzinger says. "One can shape everything."

Also appealing was the chance to return to his homeland. Henzinger hasn't lived in Austria since the 1980s, when he left to pursue a PhD at Stanford University in California.

Currently a computer and communication sciences professor at the Swiss Federal Institute of Technology in Lausanne, Henzinger has focused on developing ways to uncover bugs and errors in computer programs. He worked on this at the University of California, Berkeley, where he also co-developed a computer language that eliminates many sources of timing errors — crucial in aeroplane navigation systems, for example. If a personal computer crashes, it is an annoyance, "but if a system gets hung up in a critical moment in flight control, it could be a disaster", Henzinger says.

Although Henzinger is fascinated by the precise world of computers, he has been expanding his purview, researching applications of computer models to biology and other sciences. That multidisciplinary perspective helped make IST Austria's decision, he says. "Interdisciplinarity is a necessity in science today. At IST Austria, we will have material scientists next to biologists next to computer scientists," Henzinger says. "Computer science today is not just about number crunching but is a science of design."

Edward Lee, from the University of California, Berkeley, says that his former colleague is ideally suited to lead IST Austria: "His thinking is unconventional — he doesn't do mainstream stuff in mainstream ways. Everything he's done has branched in a new direction and had a major impact."

In his new role, Henzinger will recruit faculty members instead of being one, and he won't just fill seats. "We want to create a top-ranked research institute that will compare with the top institutes in Europe and the United States," he says. "We will not compromise." ■

Karen Kaplan

SCIENTISTS & SOCIETIES

Scientists without borders

Science job and workforce growth in the United States could be stymied under current federal controls that govern visas and exports, warns a new report, which calls for a revision of the existing regulations.

The January 2009 report by the National Research Council, *Beyond "Fortress America": National Security Controls on Science and Technology in a Globalized World*, says current visa and export regulations are rooted in the 1950s, hamper US competitiveness and impede science and technology job and industry growth.

US visa and export regulations impede the free flow of people into the country as well as information or products out of it. Recent changes in visa laws have lengthened the time it takes for a non-US resident to get a US visa, the US state department concedes on its website. Export laws limit or bar publication of information and exportation of goods that could potentially pose a threat to national security.

The regulations are driving critical jobs, and valuable discoveries and inventions, overseas, the report says.

John Hennessy, president of Stanford University in California, was co-chair of the council committee that authored the report, and Deanne Siemer, a lawyer and consultant, and Gerald Epstein, from think tank the Center for Strategic and International Studies, both based in Washington, DC, joined him on the committee. They say non-US students and scientists must have access to US universities and science labs. Research collaborations are jeopardized when non-US scientists experience delays getting a visa and can stay for only a brief period, they say.

"Increasingly, we see organizations choosing to have meetings outside the United States to avoid visa issues," Hennessy says. "If we don't permit the world's best students and scholars, scientists and technologists to come here, [science job creation] won't happen," says Epstein.

"Science and technology graduates are the ones who actually create more US science jobs," agrees Siemer.

The report recommends non-US scientists receive a visa under an accelerated skill-based selection process. "We must get that talent here faster," Siemer says. "We're talking about our economic competitiveness."

Siemer believes an executive order from the White House mandating the recommended changes could be signed. "There is a large reservoir of expertise behind this report," Siemer says. "It's likely this will be adopted." ■

Karen Kaplan

POSTDOC JOURNAL

Life's bitter-sweet symphony

An orchestra integrates disparate instrumental sections to achieve beautiful musical harmony. In the same way, for personal and professional success I seek to balance family life with the demands and pressures of a postdoc's life.

Conducting experiments, completing data analysis, writing and hearing cries of "Daddy, you're home!" are the quaint sum of my typical day. Science consumes my mind and my pipetting hand. Family consumes my heart. Striking the balance is often challenging, especially when things get hectic in the lab. Happy Valley, or so they call it, is the home of Pennsylvania State University, where I study the mechanisms governing gene regulation in baker's yeast using a genome-wide approach. As a new postdoc, I am struggling with Robert Frost's literary fork-in-the-road decision: Do I pursue a career in academia or industry?

In the sagacious words of the knight from *Indiana Jones and the Last Crusade*, "Choose wisely, but you may only choose one." In choosing between academia and industry, I see obvious advantages and disadvantages. How will I choose wisely? How will I balance my personal and professional considerations?

The key to making wise decisions is to ask advice from those who are wiser and more experienced than I am. Making life-changing decisions is not easy, but I invite you to join me over the course of the next year as I compose my own personal symphony. ■

Bryan Venters is a postdoc in molecular biology at Pennsylvania State University.

Commitment

The wheels of justice.

John Gilbey

JACEY

It was the pain that woke me, something that has happened increasingly often over the past couple of months. I thought at first that it was some residual effect from the accident, so I didn't bother my doctors with it — which was a mistake. A fatal mistake, as it turns out.

As the light spilling around the curtains got stronger, I edged slowly over to the window. Below me, Tiburon lay infinitely tranquil in the soft autumn sun. Across the Bay, framed by Angel Island on one side and the lonely towers of the Golden Gate on the other, the broken teeth of the San Francisco skyline hung in the mist. Ellen had loved this view — a love that led us to stay here even after the tsunami of 2018, the one that missed us by just a few feet. Next time I'd know what to do when the pelicans disappear — not, I reminded myself, that I will see a next time.

By the time I had showered and fought my way into my clothes, I could see that Dave was waiting for me on the street behind the house. Just the knowledge that I was going to see him again made me feel better. Don't get me wrong, it's nothing sexual, but he is by far the best assistant I have ever had — bright, intuitive and insightful — and for the past three years he has certainly been my best friend and only true confidant.

In many ways he is much like we hoped our son would be — would have been, I'm certain, if a drunk hadn't run us off the road that evening above Muir Woods. Why the safety systems failed I never found out, but we fell two hundred feet before the trees stopped us. Two died — and I have often wished it had been all three of us.

I must have looked even worse than usual that morning, for as I struggled into the seat Dave offered to drive. I shook my head.

"This could be my last chance," I said with unnecessary emphasis, thumbing the command console into life.

"I could say the same thing," he commented evenly, and I felt my usual pang of guilt.

The ferry, smugly important since we began to run out of bridges, was loading when we got to the dock. I managed a smile for the pretty young dock-master while the guidance system hauled us on board, then we were accelerated smoothly off across the Bay.

A group of early tourists watched with interest as we were disgorged onto the urban guideway in front of the Ferry Building. We slotted in behind an F-Car and headed up Market towards City Hall. Dave had been quiet during the 20-minute crossing. He usually kept me chatting about events in the lab, stuff he had picked up online, things on my to-do list that he thought I'd forgotten about — but it wasn't until we had passed 4th Street that he said: "Are you sure you want to go through with this? I'll understand if you want to change your mind."

I was surprised to find tears forming in my eyes and brushed them away impatiently. "Dave, I have no doubts about this at all. Ellen would have loved you just as much as I do — and I'm sure she would want you to inherit my estate. After all, who else am I going to leave everything to? Above all, this way I hope I can make sure that you don't get your contract revoked the moment I'm dead."

Dave hesitated a moment, no doubt thinking of the desperately unpleasant consequences of revocation, then said in confidential tones: "Thanks, but it isn't about me, you know — it can't be. Even so, if I can guard your legacy I'd be proud to be your heir. I guess all we need to do now is convince the court …"

I thought about the paperwork in my bag. The stultifyingly complex legal prose that began "In the matter of the Next Phase Institute versus the City of San Francisco …" and rambled on for some 200 pages. It all boiled down to a single concept: when does a constructed intelligence develop truly human attributes? When can an artificial person become a full member of society, complete with assets, property and social responsibilities?

We arrived at City Hall just as my legal team plodded around the corner hauling their crates of papers — hotly pursued by the news media, remote cameras floating like corporate seagulls over the throng. I tried to swing my legs out onto the sidewalk with some degree of dignity — but the pain was so intense that I had to pause for a moment.

"Good luck, John," he said, "I hope you get the outcome you want." It was the first time he had ever used my first name in public, perhaps an act calculated to let the lawyers and microphones overhear. "Thank you, Dave," I said quietly — then as an afterthought, "Are you coming in to see the fun?"

He thought for a moment. "I don't think that would be good tactics — too intrusive, I think. I may listen in online — but I'm getting pre-failure diagnostics in one of my braking systems, I really ought to go and get it checked out properly."

I patted the dash and stood up, stepping awkwardly towards my lawyers. Pausing for breath at the foot of the steps, I looked back and as Dave rolled quietly away had a moment of insight into why the case had caused so much interest around the world. But then, in California — of all places — you should be entitled in law to love your automobile. ■

John Gilbey is an author and photographer who lives and works in Wales.